

**AN INFORMATION MODELING APPROACH TO IMPROVE QUALITY OF
USER-GENERATED CONTENT**

by

© Roman Lukyanenko

A Dissertation submitted to the

School of Graduate Studies

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Faculty of Business Administration

Memorial University of Newfoundland

August 2014

St. John's Newfoundland and Labrador

ABSTRACT

Online user-generated content has the potential to become a valuable social and economic resource. In many domains – including business, science, health and politics/governance – content produced by ordinary people is seen as a way to expand the scope of information available to support decision making and analysis. To make effective use of user-generated contributions, understanding and improving information quality in this environment is important. Traditional information quality research offers limited guidance for understanding information quality issues in user-generated content. This thesis analyzes the concept of user-generated information quality, considers the limits and consequences of traditional approaches, and offers an alternative path for improving information quality. In particular, using three laboratory experiments the thesis provides empirical evidence of the negative impact of class-based conceptual modeling approaches on information accuracy. The results of the experiments demonstrate that accuracy is contingent on the classes used to model a domain and that accuracy increases when data collection is guided by classes at more generic levels. Using these generic classes, however, undermines information completeness (resulting in information loss), as they fail to capture many attributes of instances that online contributors are able to report. In view of the negative consequences of class-based conceptual modeling approaches, the thesis investigates the information quality implications of instance-based data management. To this extent this thesis proposes principles for modeling user-generated content based on individual instances rather than classes. The application of the proposed principles is demonstrated in the form of an information system artifact - a real system

designed to capture user-generated content. The principles are further evaluated in a field experiment. The results of the experiment demonstrate that an information system designed based on the proposed principles allows capturing more instances and more instances of novel classes compared with an information system designed based on traditional class-based approaches to conceptual modeling. This thesis concludes by summarizing contributions for research and practice of information/conceptual modeling, information quality and user-generated content and provides directions for future research.

ACKNOWLEDGEMENTS

To discover something new, one takes the roads less traveled. As I am reflecting on my five-year odyssey in the uncharted territories of information management, it is becoming clear that my biggest find is the many kind and caring people I met or got to know better during my intellectual travels.

It is easy to see the way when you stand on the shoulders of giants. This is what my supervisor, Jeff Parsons, has been for me. Jeff is my intellectual pillar and a true friend. I want to thank him for setting the highest standards and seeing my aspirations come true. Without Jeff I would still be walking in circles.

I am deeply indebted to my committee members, Yolanda Wiersma and Joerg Evermann for their unceasing support and insightful guidance. Without Yolanda I would have never fathomed to seek business insights in the realm of plants and animals! Thank you for taking me aboard NLNature and setting me on course to the frontiers of science. Joerg is a cornucopia of knowledge and each time I talked to him, I learned something new. Thank you, Joerg, for sharing so many valuable insights with me and always greeting my fits of curiosity with patience.

I want to thank my thesis examiners, Drs. Andrew Burton-Jones, Andrew Gemino, and Sherrie Komiak for their insightful comments and suggestions.

Family gives me the reason to work and live. I want to thank my parents, Ann and Vladimir, and my sister Victoria for their limitless love. I am forever grateful to my wife Daria for her incredible tenacity and understanding during these amazing years of intellectual pursuit. You stood by me notwithstanding the 'normal' and financially-stable

life you had before we met. I want to thank Ludmila and Katherine for their great support and friendship!

I met so many incredible people while en-route to knowledge! Ivan Saika-Voivod and his family and Stuart Durrant have been great friends during these four years. I want to thank everyone from Memorial University's Faculty of Business, School of Graduate Studies and the Harris Centre as well as The Natural Sciences and Engineering Research Council of Canada for supporting my pursuits.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS.....	IV
LIST OF TABLES	X
LIST OF FIGURES	XI
LIST OF ABBREVIATIONS	XII
1 INTRODUCTION	1
1.1 BACKGROUND AND MOTIVATION	1
1.1.1 Growth of User-generated Content.....	1
1.2 INFORMATION QUALITY CHALLENGES OF USER-GENERATED CONTENT.....	4
1.3 OBJECTIVES OF THE THESIS	7
1.4 THESIS ORGANIZATION.....	10
2 THE PROBLEM OF CROWD IQ IN EXISTING RESEARCH.....	12
2.1 DEFINING CROWD IQ.....	12
2.1.1 Traditional Views on IQ.....	12
2.1.2 IQ in UGC	14
2.2 APPROACHES TO IMPROVING CROWD IQ	17
2.3 TRADITIONAL CONCEPTUAL MODELING APPROACHES.....	23
2.4 CHAPTER CONCLUSION.....	29
3 IMPACT OF CONCEPTUAL MODELING ON INFORMATION QUALITY	30
3.1 IMPACT OF CONCEPTUAL MODELING ON DATA ACCURACY	33
3.2 IMPACT OF CONCEPTUAL MODELING ON INFORMATION LOSS	35
3.3 IMPACT OF CONCEPTUAL MODELING ON DATASET COMPLETENESS.....	36
3.4 CHAPTER CONCLUSION.....	37
4 IMPACT OF CONCEPTUAL MODELING ON ACCURACY AND INFORMATION LOSS	39

4.1	INTRODUCTION	39
4.2	EXPERIMENT 1	41
4.2.1	Impact of Conceptual Modeling on Accuracy in a Free-form Data Collection	41
4.2.2	Impact of Conceptual Modeling on Information Loss in a Free-form Data Collection	44
4.2.3	Experiment 1 Method	45
4.2.4	Experiment 1 Results	50
4.3	EXPERIMENT 2	56
4.3.1	Experiment 2 Method	58
4.3.2	Experiment 2 Results	60
4.4	EXPERIMENT 3	62
4.4.1	Experiment 3 Method	64
4.4.2	Experiment 3 Results	66
4.5	CHAPTER DISCUSSION AND CONCLUSION	70
5	PRINCIPLES FOR MODELING USER-GENERATED CONTENT	73
5.1	EMERGENT APPROACHES TO CONCEPTUAL MODELING	73
5.2	CHALLENGES OF MODELING USER-GENERATED CONTENT	77
5.3	PRINCIPLES FOR MODELING USER-GENERATED CONTENT	81
5.4	CHAPTER CONCLUSION	92
6	DEMONSTRATION OF THE PRINCIPLES FOR MODELING UGC IN A REAL CITIZEN SCIENCE INFORMATION SYSTEM.....	94
6.1	NLNATURE BACKGROUND	94
6.2	PHASE 1 DESIGN	96
6.3	PHASE 2 DESIGN	100
6.4	DISCUSSION	111
6.5	CHAPTER CONCLUSION	114
7	IMPACT OF CONCEPTUAL MODELING ON DATASET COMPLETENESS.....	116

7.1	THEORETICAL PREDICTIONS	116
7.2	METHOD	119
7.3	RESULTS	127
7.3.1	Hypothesis 4.1: Number of instances stored	130
7.3.2	Hypothesis 4.2: Number of novel species reported.	139
7.4	DISCUSSION	142
7.5	CHAPTER CONCLUSION.....	146
8	CONTRIBUTIONS, FUTURE WORK AND CONCLUSIONS	148
8.1	CONTRIBUTIONS TO RESEARCH AND PRACTICE	148
8.1.1	Reconceptualizing IQ	148
8.1.2	Exposing Class-based Approaches to Conceptual Modeling as a Factor Contributing to Poor Crowd IQ.....	149
8.1.3	Novel Approaches to Improving IQ	153
8.2	FUTURE RESEARCH.....	154
8.2.1	Impact of Conceptual Modeling on Other IQ Dimensions	154
8.2.2	Impact of Contributor-oriented IQ on Data Consumers	155
8.2.3	From UGC to Other Domains	156
8.2.4	Development of an Instance-based Conceptual Modeling Grammar	157
8.2.5	Addressing Challenges to Instance-and-attribute Approaches	158
8.2.6	Combining Instance-based Modeling with Traditional Modeling	160
8.3	THESIS CONCLUSIONS.....	161
	BIBLIOGRAPHY	163
	APPENDIX 1: IMAGES USED IN LABORATORY EXPERIMENTS IN CHAPTER 4..	182
	APPENDIX 2: SUMMARY OF OPTIONS PROVIDED IN EXPERIMENTS 2 AND 3 OF CHAPTER 4.....	187
	APPENDIX 3. ADDITIONAL ANALYSIS OF THE EXPERIMENTS 2 AND 3	191

**APPENDIX 4. SUMMARY OF THE THEORETICAL PROPOSITIONS AND
EMPIRICAL EVIDENCE OBTAINED 204**

List of Tables

TABLE 1. MAJOR CITIZEN SCIENCE PROJECTS THAT HARNESS UGC	41
TABLE 2. CHI-SQUARE (χ^2) GOODNESS-OF-FIT FOR THE NUMBER OF BASIC VS. SPECIES-GENUS LEVEL CATEGORIES	51
TABLE 3. FISHER’S EXACT TEST OF INDEPENDENCE IN CATEGORIES AND ATTRIBUTES CONDITION	52
TABLE 4. SAMPLE OF BASIC, SUB-BASIC AND OTHER ATTRIBUTES PROVIDED FOR AMERICAN ROBIN IN THE ATTRIBUTES-ONLY CONDITION	55
TABLE 5. NUMBER OF SUB-BASIC, BASIC, SUPER-BASIC AND OTHER ATTRIBUTES IN ATTRIBUTES-ONLY CONDITION	56
TABLE 6. COMPARISON OF ACCURACY IN EXPERIMENT 2: SINGLE (E2SL) VS. MULTI-LEVEL CONDITIONS (E2ML)	61
TABLE 7. ACCURACY IN SINGLE-LEVEL (E3SL) AND MULTI-LEVEL CONDITION (E3ML) FOR "FAMILIAR" SPECIES	67
TABLE 8. ACCURACY IN EXPERIMENT 3, SINGLE-LEVEL CONDITION (E3SL), MULTI-LEVEL CONDITION (E3ML) AND FREE-FORM CONDITION (E3FF)	69
TABLE 9. MODELING CHALLENGES IN UGC SETTINGS	81
TABLE 10. NUMBER OF OBSERVATIONS BY CONDITION	131
TABLE 11. EXAMPLES OF USER INPUT IN THE CLASS-BASED CONDITION THAT DID NOT FIT THE SPECIES LEVEL OF CLASSIFICATION	135
TABLE 12. EXAMPLES OF THE BASIC-LEVEL CATEGORIES PROVIDED IN THE INSTANCE-BASED CONDITION ..	137
TABLE 13. NUMBER OF OBSERVATIONS AND CATEGORIES AND ATTRIBUTES BY CONDITION	139
TABLE 14. NUMBER OF NEW SPECIES REPORTED BY CONDITION (REPEATED SIGHTINGS EXCLUDED)	140

List of Figures

FIGURE 1. THE ROADMAP AND KEY CONTRIBUTIONS OF THIS THESIS	11
FIGURE 2. INSTANCE-BASED META-MODEL	89
FIGURE 3. CONCEPTUAL MODEL FRAGMENT AND USER INTERFACE ELEMENTS BASED ON THE MODEL IN PHASE 1 NLNATURE.....	98
FIGURE 4. A VIGNETTE OF AN OBSERVATION CLASSIFIED AS MERLIN (FALCO COLUMBARIUS) WHERE THE OBSERVATION CREATOR ADMITS TO GUESSING.....	99
FIGURE 5. THE "ABOUT US" PAGE ON NLNATURE PHASE 2 THAT DESCRIBES PROJECT'S FOCUS.....	102
FIGURE 6. LOGICAL VIEW (TABLE SCHEMA) OF THE NLNATURE'S INSTANCE-BASED IMPLEMENTATION	105
FIGURE 7. EXAMPLE OF DATA COLLECTION IN PHASE II.....	108
FIGURE 8. NLNATURE PHASE 2 DATA ENTRY INTERFACE.	109
FIGURE 9. REDESIGNED FRONT PAGE OF NLNATURE (PUBLIC VIEW)	110
FIGURE 10. TRAFFIC TREND ON NLNATURE BEFORE (PRIOR TO JUNE 2013), DURING (JUNE - DECEMBER 2013) AND AFTER THE EXPERIMENT (DECEMBER 2013 TO MARCH 2014).....	121
FIGURE 11. REDESIGNED FRONT PAGE OF NLNATURE (PUBLIC VIEW)	122
FIGURE 12. CLASS-BASED DATA ENTRY INTERFACE	123
FIGURE 13. INSTANCE-BASED DATA ENTRY INTERFACE	124
FIGURE 14. NUMBER OF OBSERVATIONS PER USER IN THE TWO CONDITIONS.....	132
FIGURE 15. FEEDBACK USERS RECEIVED IN THE CLASS-BASED CONDITION WHEN THE WORD ENTERED WAS INCONGRUENT WITH THE CLASSES DEFINED IN THE MODEL.	136
FIGURE 16. A TIMELINE OF THE OBSERVATION SHOWING THE LOSS OF AN OTTER INSTANCE.....	137

List of Abbreviations

Abbreviation	Full Meaning
Crowd IQ	Crowd Information Quality
IQ	Information Quality
IS	Information System
UGC	User-generated content

1 Introduction

1.1 Background and Motivation

1.1.1 Growth of User-generated Content

Information systems (IS) were traditionally considered as being conceived, designed, implemented and used primarily within an organization for well-defined purposes determined during systems development (e.g., Mason and Mitroff 1973). This organizational focus enabled control over mechanisms to collect, store, and use data. The growth of inter-organizational systems challenged this view to some degree, as it became necessary to standardize methods for information exchange between independent systems in different organizations (Choudhury 1997; Markus et al. 2006; Vitale and Johnson 1988; Zhu and Wu 2011). The proliferation of social media (e.g., Facebook, Twitter, see Susarla et al. 2012) and crowdsourcing (engaging online users to work on specific tasks, see Doan et al. 2011) has further changed the IS landscape. There is growing interest in *user-generated content* (UGC) (Cha et al. 2007; Daugherty et al. 2008; Krumm et al. 2008), defined here as *various forms of digital information (e.g., comments, forum posts, tags, product reviews, videos, maps) produced by members of the general public – who often are casual content contributors (the crowd) – rather than by employees or others closely associated with an organization.*

Social media and crowdsourcing encourage rapid user contributions. The scale of human engagement with content-producing technologies is staggering: for example, a

2011 Pew Institute survey reports half of US adults use social media / networking websites¹. The rise of content-producing technologies offers an opportunity to collect information from anyone who has access to the Internet.

User-generated contributions increasingly support decision making and analysis in many domains. Companies nurture user-generated content by creating digital platforms for user participation (Gallaughier and Ransbotham 2010; Gangi et al. 2010; Piskorski 2011), in part to monitor what potential customers are saying (Barwise and Meehan 2010; Culnan et al. 2010). In health care, UGC promises to improve quality, for example, via feedback on hospital visits posted online (Gao et al. 2010). Many governments provide digital outlets for citizens to participate in the political process, report civic issues, or help with emergency management (Johnson and Sieber 2012; Majchrzak and More 2011; Sieber 2006). Honing in on the promises of UGC, businesses have begun to encourage employees to create and share information using internal social media and crowdsourcing platforms to augment corporate knowledge management activities (Andriole 2010; Erickson et al. 2012; Hemsley and Mason 2012).

Scientists also actively seek contributions from ordinary people, and build for this purpose novel IS that harness the enthusiasm and local knowledge of lay observers

¹<http://www.pewglobal.org/2011/12/20/global-digital-communication-texting-social-networking-popular-worldwide/>.

(citizen scientists). Citizen scientists participate in a diverse range of online projects, such as folding proteins, finding interstellar dust, classifying galaxies, deciphering ancient scripts, identifying species, and mapping the planet (Fortson et al. 2011; Goodchild 2007; Hand 2010). Citizen science promises to reduce research costs and has led to significant discoveries (Lintott et al. 2009).

Of particular interest to organizations is *structured* user-generated content (relative to less-structured forms, such as forums, blogs, or tweets). Structured user-generated information has the advantage of consistency (i.e., the form in which data is produced is known in advance), facilitating analysis and aggregation. Structured UGC can also be easily integrated into internal information systems, connecting internal processes with real-time input from distributed human sensors. Online users tend to produce vast amounts of content extremely fast (Hanna et al. 2011; Kwak et al. 2010; Susarla et al. 2012), making UGC a key contributor to "big data" or massive, rapidly growing and heterogeneous datasets (Chen et al. 2012; Heath and Bizer 2011; Lohr 2012). Structured "big UGC" enables real-time analysis and action. For example, in response to the information provided by the user, a system can automatically and immediately perform some useful action (e.g., recommend a product to buy, ask a follow-up question, flag data for verification or some follow-up action).

Organizations harnessing structured UGC can sponsor innovative information systems to address specific organizational goals or subscribe to existing general-purpose systems to supplement internal information production. For example, Cornell University launched eBird (www.ebird.com) to collect amateur bird sightings to support its

ornithology research program (Hochachka et al. 2012; Sullivan et al. 2009). The project attracts millions of bird watchers globally and, as of 2014, collects five million bird observations per month (Sheppard et al. 2014). There is also a growing cohort of general-purpose UGC applications. For instance, CitySourced (www.citysourced.com) is a US-wide project that encourages people to report civic issues (e.g., crime, public safety, environmental issues) and makes this data available to participating municipalities for analysis and action. OpenStreetMap (www.openstreetmap.org) constructs user-generated maps, thereby providing affordable geographical information to individuals, non-profit organizations and small businesses (Haklay and Weber 2008). Projects such as Amazon's Mechanical Turk (www.mturk.com) and CrowdFlower (www.crowdflower.com) maintain a virtual workforce and lease it to clients for specific projects (e.g., to classify products in an e-commerce catalog).

1.2 Information Quality Challenges of User-generated Content

Despite its pervasiveness, UGC holds potential risks. First, by opening up participation to the crowd, it is more difficult to control the content or form of data supplied. Casual users often lack domain expertise, have little stake in the success of projects, and cannot be held accountable for the quality of data they contribute (Coleman et al. 2009). To produce contributions of acceptable quality to project sponsors (e.g., scientists, e-commerce vendors, businesses or public policy makers), some level of domain knowledge (e.g., bird taxonomy, geography, consumer products) is required. However, this requirement may not generally hold for a public increasingly engaged in

content creation. As a result, there is a potential trade-off between level of participation and information quality. Ordinary people unfamiliar with the domain of a specific project may either avoid contributing or provide incorrect data (e.g., by misidentifying a bird or a product).

Second, in a crowd environment casual participants may lack incentives to contribute and may be dissuaded if the process of making contributions is difficult. For example, if an interface requires data to be recorded at a level of specificity that a casual contributor cannot easily provide, potential contributions might be lost.

Third, different contributors have different perceptions of what is relevant and interesting for a particular observation. If the system is not flexible enough to allow unanticipated data to be captured systematically, potentially useful information might be lost.

Thus, an important challenge in making effective use of UGC is *crowd information quality*² (crowd IQ) – the quality of information contributed by Internet users (Arazy and Kopak 2011; Arazy et al. 2011; Flanagin and Metzger 2008; Hochachka et al. 2012; Mackechnie et al. 2011; Nov et al. 2011a; Wiggins et al. 2011). Perceived or actual low quality of UGC can severely curtail its value in decision-making.

² Following Wang (1998) and Redman (1996), this thesis uses the terms *information* and *data* interchangeably. Crowd IQ is formally defined in Section 2.1.2.

The potential low crowd IQ poses a dilemma in harnessing collective intelligence or the “wisdom of crowds” (Surowiecki 2005). On the one hand, mounting evidence of the potential value in UGC strongly favors allowing users to freely express themselves (Hand 2010; Lintott et al. 2009). Placing restrictions on the kind of information users may wish to contribute threatens to preclude them from communicating valuable insights. On the other hand, as platforms harnessing user contributions attract more diverse audiences, restrictions upon user input seem to be necessary to ensure the quality of information collected (e.g., Hochachka et al. 2012).

Currently, there is little theoretical guidance to address emerging challenges of crowd IQ. Although information quality has been studied extensively in the information systems field, prior research focused on corporate data collection (e.g., Ballou et al. 1998; Lee 2003; Volkoff et al. 2007). A typical strategy to increasing quality in corporate environments is training of data entry operators (Redman 1996). Training or providing quality feedback appears to be considerably less effective, and often is infeasible, among casual online users. In traditional IQ management, it is considered important to ensure that all parties (e.g., data creators, data consumers) share a common understanding of what data is relevant, how to capture it and why it is important (e.g., Lee and Strong 2003, p. 33). This clearly becomes problematic in UGC settings as online users may not be willing to adopt or be capable of fully understanding the organizational perspectives.

1.3 Objectives of the Thesis

Given the limitations of traditional approaches to IQ in UGC, novel approaches are needed. This thesis examines the effect of a largely ignored, but important, factor influencing IQ in UGC – *conceptual modeling*. Conceptual modeling and IQ management have traditionally been seen as distinct activities. Conceptual modeling is concerned with representing knowledge about a domain, deliberately abstracting from implementation issues (Clarke et al. 2013; Guizzardi and Halpin 2008; Mylopoulos 1998; Wand and Weber 2002).

Conceptual modeling has been defined as “the activity of formally *describing some aspects of the physical and social world around us* for the purposes of understanding and communication” (Mylopoulos 1992; emphasis added). Conceptual models are constructed by systems analysts at the early stages of IS development to express concepts in the domain as viewed by IS users (e.g., decision makers, data consumers). Conceptual models typically inform the design of such IS artifacts as database schema, user interface, and programming code.³ By comparison, research on IQ

³ This thesis uses the term "conceptual modeling" to specifically refer to the activity of capturing concepts in the domain as viewed by data consumers (e.g., scientists) interested in harnessing UGC. Unless indicated otherwise, the resulting conceptual models are independent of implementation considerations (e.g., logical and physical representation of UGC).

has emphasized the needs of data consumers and their experiences with IQ. These experiences can be characterized using dimensions such as consistency, timeliness, believability, accessibility, security, completeness, value-added, ease of manipulation, and freedom from error (accuracy) (Lee et al. 2002; Wang and Strong 1996). Some studies suggest that the intersection of modeling and crowd IQ warrants attention. Girres and Touya (2010) note the importance of the data model used by the OpenStreetMap project, and argue for a better balance between contributor freedom and compliance with specifications.

This thesis claims that IQ is affected by decisions about underlying conceptual models. Investigating conceptual modeling as a factor affecting IQ is a promising avenue for research. Online users in UGC settings may resist traditional IQ methods such as training, instructions and quality feedback. In contrast, conceptual modeling is an activity that is typically performed before users are allowed to contribute data and thus remains firmly within organizational control. At the moment, however, little is known about the relationship between conceptual modeling approaches and crowd IQ. This thesis contributes to a better understanding of the impact of the process of creating a conceptual model of the domain on information quality. The first research question of this thesis, therefore, is:

Research Question 1: How does conceptual modeling affect IQ in UGC settings?

This thesis proposes that the IQ of structured user contributions can be positively or negatively influenced by conceptual modeling decisions. In particular, the dominant approach, in which data are conceived and recorded in terms of classes (e.g., phenomena

are assigned to classes such as *product type*, *biological species*, or *landscape form*), may have a significant negative impact on IQ when the classification structure provided by a system based on the needs of *data consumers* (e.g., decision-makers in the organization looking to draw insights from UGC) does not align with that of *data contributors* (i.e., the online users participating in UGC projects and contributing data). Once defined, classes constrain the degree to which an information system is able to reflect users' views of reality. Relaxing the rigid constraints of *class-based* models may help in capturing user input more objectively and completely, leading to higher quality of stored data while simultaneously mitigating the constraints on participation arising from insufficient expertise and differences in domain conceptualizations among online users. It may also fuel discovery by creating an environment that facilitates the discovery of previously unknown classes of phenomena. This further promises an opportunity to use conceptual modeling as a mechanism for crowd improving IQ. Therefore, the second research question is:

Research Question 2: What conceptual modeling principles can be developed to improve quality of UGC?

As traditional modeling approaches may have detrimental effects on crowd IQ, the thesis raises the question of what alternative approaches may help mitigate the shortcomings of traditional modeling. The thesis thus proposes theory-based principles for modeling UGC, intended to improve crowd IQ while relaxing restrictions on the kind of information users can provide.

1.4 Thesis Organization

The remainder of the thesis is organized as follows (see also Figure 1). The next chapter situates the problem of crowd IQ in the context of the current conceptualizations of IQ and conceptual modeling. As the chapter uncovers the limitations of the prevailing approaches to IQ in UGC settings, it proposes a novel definition of crowd IQ.

Chapter 3 provides a theoretical foundation for crowd IQ and conceptual modeling and uses theories in philosophy and psychology to derive propositions about the impact of conceptual modeling on important IQ dimensions of accuracy and completeness (including information loss and dataset completeness).

Chapter 4 presents three laboratory experiments that test hypotheses about the impact of conceptual modeling on *accuracy* and *information loss* based on the propositions from Chapter 3.

Chapter 5 develops principles for modeling UGC intended to address identified challenges of IS development in these settings.

Chapter 6 demonstrates how to model UGC following the principles proposed in Chapter 5 in the form of an information system artifact - a real system designed to capture UGC.

Chapter 7 presents a field experiment in the context of citizen science in biology and evaluates the impact of conceptual modeling approaches on *dataset completeness*.

The thesis concludes by summarizing the primary contributions of the research to theory and practice and suggesting several areas for future research.

Overall objective: Improving Information Quality in User-generated Content	
<div> Research Question 1: How does conceptual modeling affect information quality in UGC settings? </div>	
Chapters 1, 2	<ul style="list-style-type: none"> • Problem of managing information quality of UGC • Definition of crowd IQ • Limitations of existing approaches to crowd IQ • Identification of conceptual modeling as a promising direction • Exposition of the gap in understanding how conceptual modeling affects crowd IQ
Chapter 3	<ul style="list-style-type: none"> • Theoretical explanation of the potential impact of conceptual modeling on <ul style="list-style-type: none"> ◦ information accuracy ◦ information loss ◦ dataset completeness
Chapters 4 and 7	<ul style="list-style-type: none"> • Three laboratory experiments to evaluate the impact of conceptual modeling on: <ul style="list-style-type: none"> ◦ accuracy ◦ information loss • Field experiment to evaluate the impact of conceptual modeling on: <ul style="list-style-type: none"> ◦ dataset completeness • Summary of findings: <ul style="list-style-type: none"> ◦ Traditional approaches to conceptual modeling may have negative impact on accuracy, information loss and dataset completeness dimensions of IQ
<div> Research Question 2: What conceptual modeling principles can be developed to improve quality of UGC? </div>	
Chapter 5, 6 and	<ul style="list-style-type: none"> • Principles of modeling UGC based on representation of instances (rather than classes) • Demonstration of the proposed principles in the form of a real IS • Evaluation of the proposed principles in a field experiment <ul style="list-style-type: none"> ◦ IS designed based on the proposed principles allows capturing more instances and more instances of novel classes compared with IS designed based on traditional approaches to conceptual modeling
Chap	<ul style="list-style-type: none"> • Thesis contributions • Directions for future research

Figure 1. The roadmap and key contributions of this thesis

2 The Problem of Crowd IQ in Existing Research

2.1 Defining Crowd IQ

2.1.1 Traditional Views on IQ

Information quality has been studied extensively in the information systems field, with the primary focus on corporate uses of IS, in which user input may be relatively well-controlled (Ballou et al. 1998; Madnick et al. 2009; Storey et al. 2012; Wang and Strong 1996). In this environment, it is common to distinguish three parties to IQ processes: users who create data, IT professionals who secure, maintain and store it, and data consumers (Lee 2003). These three parties are typically in close contact and work jointly to refine and improve information quality (e.g., IT professionals may coach data entry operators; data consumers may monitor and evaluate information quality). The *context* (Lee 2003) in which information was produced, managed and used was frequently amenable to scrutiny and change (for a review of IQ research, see Madnick et al. 2009).

A core principle of traditional IS analysis and design is user-driven development, according to which user (or, more commonly, eventual data consumer) requirements are captured during systems analysis and reflected to the extent possible in the design of the resulting information system (Checkland and Holwell 1998; Hirschheim et al. 1995). This consumer-oriented view is reflected in seminal definitions of information quality: the prevailing conceptualization of IQ is *fitness for use* of data by information consumers for specific purposes (Lee and Baskerville 2003; Lee and Strong 2003; Wang and Strong 1996; Zhu and Wu 2011). This focus underlies another popular IQ definition –

“conformance to specification and as exceeding consumer expectations” (Kahn et al. 2002). Both definitions focus IQ improvement on ways to shape the “information product” (Ballou and Pazer 1985; Wang 1998) to better satisfy data consumers’ needs and are concomitant with conceptions of quality in marketing and management science (Juran and Gryna 1988; Reeves and Bednar 1994).

The conceptualizations of dimensions of IQ further adopted the fitness for use perspective. Thus, Parssian et al. (2004) define completeness "as availability of all relevant data to satisfy the user requirement" (p. 968). Lee et al. (2002) developed measurement items to evaluate completeness, asking whether "information includes all necessary values", "information is sufficiently complete for our needs", "information covers the needs of our tasks", "information has sufficient breadth and depth for our needs" (p. 143). To this extent, completeness has been classified as a contextual IQ dimension (Wang and Strong 1996). Nelson et al. (2005) explain (p. 203):

It is important to recognize that the assessment of completeness only can be made relative to the contextual demands of the user and that the system may be complete as far as one user is concerned, but incomplete in the eyes of another. While completeness is a design objective, its assessment is based on the collective experience and perceptions of the system users.

In consumer-focused IQ, it becomes important to ensure that all parties to IQ management (e.g., data creators, data consumers) share a common understanding of what data is relevant, how to capture it and why it is important; Lee and Strong describe this process (2003, p. 33):

To process organizational data, a firm’s data production process is conceptually divided into three distinct areas: data collection, data storage, and data utilization. Members in each process, regardless of one’s functional

specialty, focus on collecting, storing, or utilizing data. To achieve high data quality, all three processes must work properly.

Most organizations handle data quality problems by establishing routine control procedures in organizational databases. To solve data quality problems effectively, the members in all three processes must hold and use sufficient knowledge about solving data quality problems appropriate for their process domains. At minimum, data collectors must know what, how, and why to collect the data; data custodians must know what, how, and why to store the data; and data consumers must know what, how, and why to use the data.

2.1.2 IQ in UGC

Important differences between traditional organizational settings and UGC applications require extending the prevailing data consumer focus of IQ. Consumer-centric definitions ignore the characteristics of crowd (volitional) information creation and may not reflect the information contributor's perspective. UGC projects are often designed at the request of project sponsors – those who allocate resources (e.g., financial, management, and technical) to the project and evaluate its success in serving the needs of (potential) data consumers. However, ordinary people are the key contributors of information and the main drivers of success in these projects. The abilities, motivation, and domain knowledge of contributors in UGC can have a strong impact on the level of engagement and quality of contributions (Coleman et al. 2009; Hand 2010; Nov et al. 2011b). Furthermore, contributors to UGC projects may be neither aware of the intended use of contributed data nor motivated to fully satisfy (or exceed) expectations of data consumers (Daugherty et al. 2008; Nov et al. 2011a; Nov et al. 2011b). Overemphasizing the data consumer's perspective in systems designed to harness UGC may preclude

contributors from accurately and fully describing the phenomena about which they are contributing data. In cases where the data consumer's information needs are incongruent with what a user can provide, potential contributors may simply abandon data entry. Often contributors provide what they are able (or are willing), not necessarily what is required. Such information can be useful for purposes not anticipated when a project was designed. To be effective, information systems in UGC settings should be sensitive to information contributors' capabilities, as well as to data consumers' requirements.

In an online environment, traditional processes of quality control break down. Reaching and influencing (e.g., training, providing quality feedback to) content creators is often infeasible. The role of information producers and consumers is frequently blurred, making it difficult for information consumers to evaluate the quality of their own contributions. Finally, the context of information production (and, rarely, information consumption) is opaque (e.g., the conditions under which online contributors make observations may drastically vary). The nature of crowd information precludes a straightforward application of traditional principles of information quality management.

The thesis therefore proposes a definition of crowd IQ that amends the traditional definition of information quality to account for the issues and challenges of the emerging area of UGC. Specifically, ***crowd Information Quality (crowd IQ)*** is defined as *the extent to which stored information represents the phenomena of interest to data consumers (and project sponsors), as perceived by information contributors*. This definition does not rely on “fitness for use”, but is driven by what data contributors consider relevant when they use an IS. It is use-agnostic, recognizing that “the

phenomena...as perceived by information contributors” accommodates both known uses and future, unanticipated uses.

A consequence of a use-agnostic notion of IQ is that *information relevance* is “irrelevant,” as relevance must be evaluated with respect to some use or purpose. Data provided by online contributors may be collected with one use in mind (and may not be relevant for that use), but used for many different tasks and support anticipated future uses.

Crowd IQ assumes that any information about some “phenomena of potential interest” to data consumers is better than (or no worse than) no information at all, as information irrelevant to a particular use can be ignored/filtered (e.g., a query on species observed in some area will ignore contributions that are not reported at the species level).

At the same time, the definition is explicitly concerned with the needs of data consumers - who typically sponsor or have other vested interests in the success of UGC projects. Thus, UGC quality is evaluated and measured by data consumers. For example, a contributor to a citizen science project in biology (e.g., eBird.org) may classify a bird as *American robin*. The extent to which this is accurate (in this case accords with the established biological nomenclature) is left up to the data consumers (e.g., scientists) to determine (assuming they have an independent way to verify the observation). As demonstrated in more details in Chapters 4 and 7, this thesis allows the contributors to determine what information to provide, which results in higher information accuracy and completeness (as measured by data consumers).

The Crowd IQ definition provides guidance for research aimed at improving the quality of UGC. By addressing consumer needs, this thesis advocates making IQ improvements that lead to desirable and useful outcomes for consumers. At the same time, the definition recognizes the pivotal role of information contributors and motivates an effort to design systems sensitive to their points of view.

2.2 Approaches to Improving Crowd IQ

In response to the growing interest in UGC, two perspectives on how to better understand and improve crowd IQ have emerged. Consistent with broader IQ research, the prevailing approach is *fitness for use*, which focuses on the organization, qualifications and expertise of contributors so as to better align information capture with needs of data consumers. This approach assumes that potential uses of information are known and understood by data contributors (in contrast, the thesis advocates a *contributor-oriented* perspective that examines ways to design IS to better capture observations of information providers). Below I briefly consider some of the emerging approaches to crowd IQ.

Considering low domain expertise of users to be the principal detriment to high information quality, some research investigates the role of organizational processes governing information collection on data quality. Here a central element of social media, collaboration among users, is considered important. For example, this approach is the basis for iSpot (www.ispot.org.uk), a project that relies on social networking for collaborative identification of species of plants and animals (Silvertown 2010). Collaboration is also at the heart of Wikipedia (Arazy et al. 2011). The success of the

iterative process by which Wikipedia articles are refined suggests that data quality may, in fact, improve with continuous use. Social networking is suggested to increase data quality through the increased scale of data collection. According to Heipke (2010), in crowdsourcing “from a statistical point of view one can expect to have a rather low rate and size of errors” (p. 553).

While peer or collaborative review appears promising, it has a number of limitations. Despite being likened to the “scientific peer review process” (Bishr and Mantelas 2008, p. 235), peer review is appropriate only for projects with a large number of users. Web sites with a small number of users will not have sufficient user activity per unit of data to ensure adequate critique, but even in larger projects less popular content may escape peer scrutiny (Cha et al. 2007). The peer review process also raises a philosophical issue of whose perceived reality is being represented and stored: that of the original user who submitted data or that of other users who verified and corrected it? Finally, extensive collaboration often engenders task-related conflicts among members, which can diminish the quality of the product unless conflict-mediating mechanisms are in place (Arazy et al. 2011).

Another measure is engineering online governance structures (e.g., hierarchies of users), in which contributions are constrained by the organizational roles of their authors. For example, in order to edit certain content of Wikipedia or OpenStreetMap, one needs to have moderating or administrative privileges. Ensuring high quality on Wikipedia requires an elaborate and complex system of coordination. The basic assumption underlying this approach is that users in different roles (e.g., moderator vs. rookie

member) tend to produce information that differs in quality. Arazy et al. (2011) demonstrated the importance of content-oriented members as sources of domain expertise, and administrative members as mediators of internal conflicts. Liu and Ram (2011) found that users engaging in different collaboration patterns on Wikipedia (e.g., moderation, editing, and new content production) tend to produce data that differs in quality. Despite the benefits, user specialization and structures that support it have a propensity to create what Kittur et al. (2007) call the online “elite” or “bourgeoisie,” wherein a few privileged users control the collaborative enterprise. In extreme cases, this may lead to information censorship.

Considering quality to be rooted in expertise, organizations attempt to educate and train users. Here, intensive user interaction and training are frequently prescribed. Intensive interaction among users tends to foster learning and domain expertise. Most collaborative projects benefit from users supporting and educating each other.

Quality improvement via user interaction is a passive strategy. Training, on the other hand, is an active process enacted by project sponsors. It is typical in domains with high demands for data quality and established standards to which contributions should adhere (Dickinson et al. 2010; Foster-Smith and Evans 2003). For example, in Galaxy Zoo (www.galaxyzoo.com), users are required to pass a tutorial before they are allowed to classify galaxies (Fortson et al. 2011). However, training can sometimes introduce biases as participants who know the objective of the project may overinflate or exaggerate information (Galloway et al. 2006). In addition, training is not always realistic, especially among uncommitted online users. Some training requires gradual acquisition of

knowledge over time, which can be prohibitive among casual contributors. Finally, depending upon the scope of a project, the knowledge gap might be too large to bridge in a short span of time (e.g., iSpot accepts observations of all natural history phenomena, and Wikipedia allows users to contribute to any article).

Quality can also be enhanced after data is produced. Content filtering is a form of design-oriented data quality that aims to maximize the quality data of a given data set (e.g., by verifying it or only considering contributions matching certain criteria). Here, there may be no contributor manipulation before data entry, as data can be collected “as is” and filtered to retrieve only that of acceptable quality. Filtering may be performed by experts, peers or intelligent artificial agents. For example, eBird uses a combination of human and machine verification mechanisms to filter bird sightings (Hochachka et al. 2012; Sullivan et al. 2009). Content filtering (or data cleaning) typically precedes more complex analysis of UGC (Provost and Fawcett 2013).

As the size of data sets increases, manual verification becomes less realistic (e.g., Delort et al. 2011; Hochachka et al. 2012). Verification is also impossible for evanescent events that are over before experts can verify observation accuracy (e.g., vagrant bird sightings). At the same time, it can be difficult to develop automatic procedures that can deal with the full range of unanticipated UGC. Data filtering for some crowdsourcing projects, such as the website www.oldweather.org, where users transcribe historical ship logs, can only be verified by cross-validation between peers, since the task at hand (interpreting hand writing) requires human cognitive skills and is not something a computer can readily be trained to do. As with peer verification, content filtering raises

concerns about the final data reflecting biases and perceptions of humans or agents involved in the verification process.

In contrast to the use-oriented approaches to crowd IQ, this thesis investigates ways to design IS to better capture observations of information providers. Specifically, this thesis proposes conceptual modeling as a mechanism for improving crowd IQ. Investigating conceptual modeling as a factor affecting IQ appears promising. Online users in the UGC settings may resist traditional IQ methods such as training, instructions and quality feedback. In contrast, conceptual modeling is an activity that is typically performed before users are allowed to contribute data and thus remains firmly within organizational control.

Currently, there is little research on the impact of conceptual modeling on information quality. The connection between conceptual modeling and information quality is not well understood. This may be partially due to the fact that conceptual modeling and information quality management are generally seen as distinct activities. Conceptual modeling is concerned with representing knowledge about a domain, often deliberately abstracting from implementation concerns (Mylopoulos 1998; Olivé 2007; Wand and Weber 2002), while research on information quality typically examines dimensions of quality in existing databases (Arazy and Kopak 2011; Tayi and Ballou 1998; Wang and Strong 1996).

It is further unclear how to carry out conceptual modeling of UGC. Modeling UGC appears to be significantly different from modeling corporate domains, since reaching all potential (and even all representative) online users and reconciling their

views may not be feasible. Finally, information quality has been generally outside the scope of conceptual modeling research that has been traditionally more concerned with more proximal consequents such as the ability of users to comprehend and verify conceptual models (Bodart et al. 2001; Burton-Jones and Weber 1999; Burton-Jones and Meso 2006; Burton-Jones and Meso 2008; Figl and Derntl 2011; Gemino and Wand 2005; Parsons and Cole 2005; Parsons 2011; Recker et al. 2011; Topi and Ramesh 2002).

In a study of data quality in OpenStreetMap, Girres and Touya (2010) note the importance of the database model used by the project and argue for a better balance between contributor freedom and compliance to specifications. In a seminal theoretical article on IQ, Wand and Wang (1996) draw upon ontological theory to examine the extent to which an IS permits mapping of lawful states of reality to states of the IS. Wand and Wang, however, do not specifically consider conceptual modeling grammars or methods.

This thesis aims to increase theoretical understanding of the impact of conceptual modeling on information quality. Underlying the prevailing conceptualization of IQ is the assumption that quality depends on the contributor's expertise. Since only a small number of *potential* contributors are experts, this implies that the best data quality can come from a limited number of people. Such an approach can thereby severely limit the scope of UGC. Furthermore, the focus on expertise assumes a *particular intended use* of collaborative data (i.e., expertise in *something*). Yet, harnessing the "wisdom in crowds" presents an opportunity to embrace diverse and unanticipated insights and uses of information. Recognizing UGC as a source of unanticipated insights, some scientists are considering the benefits of collecting citizen data in a hypothesis-free manner (Wiersma

2010). In this context, I aim to develop an information quality approach that does not depend on user expertise or intended use.

2.3 Traditional Conceptual Modeling Approaches

Concomitant with traditional research on IQ, traditional approaches to conceptual modeling generally assumed corporate settings. Major tenets of traditional conceptual modeling research included user-, use- and consensus-driven development, whereby users of information (stakeholders, subject-matter experts) specify intended functions of the system and provide supporting requirements. This perspective, therefore agrees with the *fitness for use* paradigm of traditional IQ research (Lee 2006; Lee 2003; Strong et al. 1997; Wang and Strong 1996). Below I briefly examine key assumptions of traditional conceptual modeling research that I argue are problematic in UGC settings.

A core principle of traditional modeling is design in anticipation of typical uses of an IS. For example, UML diagrams typically originate in use cases that communicate at a high level the purposes for the designed system including data flows and activities to support (Jacobson et al. 1999). Once the system is designed, its quality is assessed insofar as it provides functionality and information necessary to fulfill the needs of its users (DeLone and McLean 1992; Petter et al. 2013). The uses and purposes of the IS originate in users and are determined at the earliest stages of development.

Traditionally, analysts rely on users (or, more generally, stakeholders) for subject-matter expertise and system requirements. The information is typically elicited through direct contact with end-users or their representatives (e.g., supervisors, team leaders). Analysts are thus freed from having to become domain experts and are mostly proscribed

from relying on their own independent judgment about modeled domains: “[i]n general, assumptions are made by the problem owners” (Kotiadis and Robinson 2008, p. 952). Similarly, research on conceptual modeling grammars assumes user views as given, however derived or “impoverished” they may be (e.g., Wand and Weber 1995, p. 206). At the same time, cognitive models and biases of users have been investigated with the objective of increasing the veracity of users’ assumptions about domains (Appan and Browne 2012; Appan and Browne 2010; Browne and Ramesh 2002). As users provide information requirements, it becomes vital to ensure that all representative users have been considered during requirements determination.

The availability of users made it possible for analysts to gather requirements, verify their fidelity, and resolve any conflicting perspectives before implementation (Dobing and Parsons 2006; Gemino and Wand 2004). As users were mostly employees or parties closely affiliated with the organization (e.g., clients, suppliers, business partners), any individual or divergent views were generally subsumed by an agreed-on view. Existing organizational structures made it easier for analysts to discover user perspectives and resolve any conflicts. Close contact with users, such as in joint or participative development is widely encouraged (Gould and Lewis 1985; Moody 2005; Mylopoulos 1998). In contrast, “lack of user input” is considered among “leading reasons for project failures” (Gemino and Wand 2004, p. 248).

Given the centrality of users to information systems development, analysts are encouraged to be directly engaged with users. Gould and Lewis (1985), for example, stipulate “bringing the design team into *direct contact* with potential users, as

opposed to hearing or reading about them through human intermediaries, or through an ‘examination of user profiles’” (p. 301, original emphasis). Indeed, an important role of conceptual models is facilitating mutual understanding and supporting user-analyst communications (Wand and Weber 2002).

Traditional corporate environments made it feasible to strive for *complete* and *accurate* requirements (Olivé 2007; Wand and Weber 2002), provided that an adequate elicitation process that mitigates biases takes place (Appan and Browne 2012). With much research and practice premised on having accurate and complete information available as input to conceptual modeling, scant attention has been paid to modeling when all representative users are not available.

A final conceptual model typically represents a global, integrated view of a domain but often does not represent any view of an individual user (Parsons 2003). Close contact with users provides an opportunity to resolve conflicts in individual views and generates an agreed-upon conceptualization of a domain: “[t]he difficulty here lies in conflict identification (how to find out that there is a conflict), rather than in conflict resolution (usually, one view is modified to remove the naming conflict)” (Spaccapietra and Parent 1994, p. 259-260). Analysts thus turn to relevant stakeholders to determine how to resolve conflicts: “conflict must be solved through communication among people” (Pohl 1994, p. 250). This parallels a typical organizational process of reaching a collective judgment through dialog, negotiation or specialized techniques (Easterby-Smith et al. 2012; Eden and Ackermann 1998). The unified global schema then serves as

“the basis for understanding by all users and applications” (Roussopoulos and Karagiannis 2009).

The fundamental approach to conveying domain semantics in a unified conceptual model is representation by abstraction (Mylopoulos 1998; Peckham and Maryanski 1988; Smith and Smith 1977). Abstraction enables analysts to deliberately ignore the many individual differences among phenomena and represent only *relevant* information, where consumers of data determine what is relevant. Abstraction is foundational to major conceptual modeling grammars. For example, a typical script made using the popular entity-relationship (ER) or Unified Modeling Language (UML) grammars may depict classes (which are similar to kinds, entity types, categories), attributes of classes (or properties) and relationships between classes. Classes (e.g., *student*, *tree*, *chair*) abstract from differences among instances (e.g., a *particular student*, or a *specific chair*), instead capturing the perceived equivalence of instances. Indeed, many conceptual modeling grammars consider instances (objects) to be members of their classes (entity types): “[o]ne principle of conceptual modeling is that domain objects are instances of entity types” (Olivé 2007, p. 383). Abstraction-based modeling is critical to “organize the information base and guide its use, making it easier to update or search it” (Mylopoulos and Borgida 2006, p. 35). With representation by abstraction as a modeling method, it is then possible to *completely* and *accurately* represent *relevant* domain semantics: “a conceptual schema is the definition of the general domain knowledge that the information system needs to perform its functions; therefore, the conceptual schema must include *all* the required knowledge” (Olivé 2007, p. 29, emphasis added).

The goal of accurate and complete specifications (for intended uses) has been the cornerstone of conceptual modeling since the early days (e.g., Parnas 1972) and persists to this day (Burton-Jones et al. 2013; Lukyanenko and Parsons 2013). At the same time, challenges and limitations of conceptual modeling have been well-researched. One challenge is effectively engaging subject-matter experts to identify and record relevant information (Appan and Browne 2010; Browne and Parsons 2012). Another is to ensure that grammars are expressive enough to capture the semantics important to the users (Clarke et al. 2013; Wand and Weber 1993). To ensure that users can then verify the captured semantics, conceptual models further require clarity and understandability (Bodart et al. 2001; Gemino and Wand 2005; Topi and Ramesh 2002). Wand and Wang (1996) note inherent limitations of traditional modeling in capturing unanticipated information. The notion of “complete and correct set of requirements” that “sweeps away the multiple perspectives and ambiguities of organizational life” has been criticized by interpretive researchers (Walsham 1993, p. 29). The challenges of view integration arising as a result of traditional modeling assumptions have been explored (Parsons and Wand 2000; Parsons 2003). Parsons and Wand (2000) examined the negative consequences of inherent classification (a major form of abstraction) on conceptual modeling and database operations. Samuel (2012) argues that abstraction-driven grammars impose cognitive effort by forcing users to identify instances that fit the predefined abstractions. Reaching remote users, especially on the Internet, has also been noted as a modeling challenge (Wand and Weber 2002). Despite these shortcomings,

traditional approaches to conceptual modeling continue to dominate and are also being adopted in UGC (e.g., Wiggins et al. 2013).

This survey of traditional conceptual modeling research suggests a number of reasons why employing these approaches to modeling UGC may be problematic. In contrast to more traditional settings where information creation was (or was assumed to be) well understood and controlled, in UGC projects there are typically no constraints on who can contribute information. Indeed, engaging broad and diverse audiences is their *raison d'être*. While traditional systems represented a "consensus view" among various parties, the diverse and often unpredictable user views in UGC settings makes it infeasible to reach such consensus. Finally, whereas more traditional systems supported predefined uses of data, in opening IS to the external environments, organizations hope to discover something new, triggering flexible and innovative ways to use and re-use collected information.

When developing conceptual models for UGC, some requirements may originate from *system owners or sponsors* - a relatively well understood group - but the actual information comes from distributed heterogeneous users. Many such users *lack domain expertise* (e.g., product taxonomy or deep medical knowledge) and have unique views or conceptualizations that may be incongruent with those of project sponsors and other users (Erickson et al. 2012). Unable to reach every potential contributor, analysts may not be able to construct an accurate and complete representation of modeled domains. I argue that fundamental assumptions about modeling may not hold in UGC environments and modeling using traditional grammars may result in poor IQ. The next chapter uses

theories of ontology and cognition to derive specific propositions about the impact of conceptual modeling on crowd IQ.

2.4 Chapter Conclusion

This chapter reviewed existing research in IQ and conceptual modeling as it relates to UGC. Previous research on IQ paid relatively scant attention to factors related to data contributors and focused instead on satisfying data consumers' needs. In contrast, this chapter argued IS in UGC settings should be sensitive to information contributors' capabilities, as well as to data consumers' requirements. This chapter proposed a definition of crowd IQ that amended the traditional definition of information quality to account for the important role of information contributors in UGC. It then identified conceptual modeling as a promising mechanism for improving crowd IQ.

A survey of conceptual modeling research, however, revealed inadequacies of existing approaches to modeling UGC. In contrast to more traditional settings where information creation was (or was assumed to be) well understood and controlled, in UGC there are typically no constraints on who can contribute information and engaging broad and diverse audiences is highly desirable. Applying traditional modeling to UGC environments may result in poor IQ. Chapter 3 proposes specific mechanisms by which conceptual modeling affects quality.

3 Impact of Conceptual Modeling on Information Quality

As implied by the proposed definition of crowd IQ, stored information should, to the extent possible, reflect the views of data contributors. Having identified conceptual modeling as a promising factor for improving IQ in the previous chapter, this chapter investigates the impact of class-based conceptual modeling on IQ. Specifically, I draw on theories of ontology and cognition to propose specific mechanisms by which conceptual modeling affect quality. As conceptual modeling deals with representing the world as understood by humans (Hirschheim et al. 1995; Wand et al. 1995), two theoretical foundations have been shown to be appropriate for understanding conceptual modeling grammars – ontology and cognition.

Ontology, the philosophical study of what exists, has been used as a theoretical foundation of conceptual modeling to prescribe modeling constructs and evaluate the fidelity with which models represent reality (Guizzardi 2010; Wand and Weber 2002; Wand et al. 1995). Bunge's (1977) ontology has been popular in conceptual modeling research as it maps well to IS constructs (Wand and Weber 1990) and has been able to explain and predict a variety of information systems phenomena (Burton-Jones and Meso 2006; Gemino and Wand 2005; Indulska et al. 2011; Shanks et al. 2008; Weber 1996). It has also been used to theoretically derive data quality dimensions (Wand and Wang 1996).

As human understanding of the real world is moderated by cognitive processes, it is appropriate to augment ontology with theories of cognition. In particular, classification theory “attempts to explain the nature of concepts (categories/classes) and why humans

classify” phenomena (Parsons 1996, p. 1438). Importantly, prominent conceptual modeling grammars, such as the Entity-Relationship (ER) model and Unified Modeling Language (UML) Class Diagrams, rely on class constructs (e.g., ER entity types, UML classes). Based on these foundations, I evaluate prevailing approaches to conceptual modeling and examine the potential impact of conceptual modeling on IQ.

According to Bunge, the world is made of “things” (individuals or entities). Every thing possesses properties; properties do not exist independent of things. People are unable to directly observe properties, and see them instead as attributes. Properties of things may change over time.

Things possessing common properties can be grouped together to form kinds (which are similar to classes). Unlike material things, classes (kinds) exist in human minds (Parsons and Wand 2008). According to cognitive theories, classes provide *cognitive economy* and *inference*, enabling humans to efficiently store and retrieve information about phenomena of interest (instances) (Parsons 1996; Posner 1993; Rosch and Muller 1978). In particular, cognitive economy is achieved by focusing on shared attributes, ignoring differences among instances deemed irrelevant in a particular situation.

The notion of class is a core conceptual modeling construct (Parsons and Wand 2008). Indeed, the prevailing method of representing information in an IS is recording an instance in terms of usually one *a priori* defined class (cf. Parsons and Wand 2000). This means instance information in a database derived from a class-based conceptual model is constrained by the properties of the classes to which the instance belongs. For example,

Tsichritzis and Lochovsky (1982) define *datum* (data item) in a strictly-typed data model as members of an *a priori* class. Therefore “data that do not fall into a [class]... have either to be subverted to fall into one, or they cannot be handled in the data model” (Tsichritzis and Lochovsky 1982, p. 8). Information about an instance that is not captured in any class to which it belongs cannot be captured in a class-based conceptual model or in a database designed from it (Parsons and Wand 1997).

This thesis examines the impact of storing instances in classes on two key IQ dimensions – accuracy and completeness. While research recognizes more than a dozen IQ dimensions (Wand and Wang 1996), accuracy and completeness are the most heavily studied (Redman 1996; Wand 1996). In this thesis, information completeness is broken down into two dimensions: dataset completeness (that is concerned with the number of instances stored) and information loss (or the extent to which perceived attributes of instances are captured).

First, there is a potential mismatch between the classes familiar to a contributor and those defined in the IS. A class is a mental model of perceived reality learned or derived from prior experience (Murphy 2004). Thus, a contributor may reasonably see an instance as a member of a different class than the one(s) defined for an IS. When required to conform to the class structure imposed by an IS, a contributor may classify an observed phenomenon incorrectly (from the data consumer perspective, as follows from the proposed definition of crowd IQ), leading to lower data *accuracy* (i.e., whether a statement $C(x)$ about an instance, x 's, membership in class C is true or false). For example, a system may provide classes C_1, \dots, C_N , while a contributor may see an

observation as a member of class Y (Y may be more general than any of C_1, \dots, C_N , or orthogonal to that structure). If the contributor is forced to guess (C_i), the statement $C_i(x)$ may be false, but if s/he can classify the observation confidently as an instance of Y , the statement $Y(x)$ will be true.

Second, class-based models may have a negative effect on data *completeness* (i.e., the degree to which observed information about an instance is captured). Class-based models inevitably result in property loss, as no class is able to capture all potentially observable properties of an instance. Ontologically, every “thing” is unique by the virtue of having unique properties: “what makes a thing what it is, i.e., a distinct individual, is the totality of its properties: different individuals fail to share some of their properties” (Bunge 1977, p. 111). Classification is based on similarity (shared properties) of instances and ignores properties deemed irrelevant for the purpose of classification. Therefore, completeness is necessarily reduced whenever a class is used to store instances. Below I elaborate on this analysis and develop two theoretical propositions regarding accuracy and completeness.

3.1 Impact of Conceptual Modeling on Data Accuracy

Accuracy is frequently suggested as the closest proxy for IQ (Ballou and Pazer 1995; Wand 1996; Wand and Wang 1996). Accuracy is typically defined as degree of conformity of a stored value to the actual (reference) value (Ballou and Pazer 1995; Pipino et al. 2002; Redman 1996; Wand 1996), or to some accepted fact in a domain (e.g., Barack Obama was born August 4, 1961).

As classes are observer-dependent, differences in prior experience, domain expertise, or intended uses may result in the same thing being classified differently by different people and by the same person over time (Barsalou 1983; McCloskey and Glucksberg 1978; Murphy 2004). For example a *passport* can be an *identity document*, a *thing to take on a trip abroad* and an *item to take from a burning house* (see Barsalou 1983). Naturally, humans employ only those classes with which they are familiar. People also attempt to match candidate classes to the situation at hand (Winograd and Flores 1987). Thus, the process of classification is a fluid interplay of context, purpose and prior knowledge. In contrast, class-based models *require* information contributors to conform to a particular classification (presumably driven by some predefined uses of data). In general, we assume that in the context of UGC it is impractical to determine the set of classes that would be familiar and natural to use for each potential contributor in every situation. If the set of classes presented by the system is unfamiliar to an information contributor or is incongruent with a contributor's domain conceptualization, the result may be a forced choice that does not reflect reality as perceived by the contributor and may be inaccurate with respect to a reference value adopted by the data consumers (e.g., the species of bird selected by a non-expert contributor to a system that classifies bird sightings may not be biologically correct).

Proposition 1 (Classification Accuracy): Class-based conceptual models result in lower information accuracy (more classification errors) when the classes defined in an information system do not match those familiar to the information contributor.

3.2 Impact of Conceptual Modeling on Information Loss

Support for the classification accuracy proposition would suggest the potential benefit of implementing IS that employ classes more familiar to potential contributors (assuming they could be determined in advance). While this can increase classification accuracy, it will fail to prevent a second problem – information (property) loss.

Using classes to store information about instances will always result in a failure to fully capture reality, no matter how “good” the chosen classes are. According to Bunge, any complex instance has a large number of attributes and no one class can encompass them all. Here lies a key difference between *human* and *computerized* representation. When humans classify, they *focus* on some equivalence among instances, but remain aware of individual differences. In contrast, when instances are stored only as members of classes derived from class-based conceptual models, attributes not captured by class definitions are lost. For example, if one defines a class *student* (assuming it has no subclasses) in an IS, every instance of that class will possess *only* those attributes that are part of the class definition. All other attributes will be lost. However, a human encountering a particular *student* may easily notice additional attributes of the individual (e.g., works part-time) that are not implied by the fact the person is a *student*, even if *student* is the class the person initially associates with that instance. As (ontologically) classes are unable to capture all instance attributes that might be observed, class-based conceptual models will result in information loss as long as contributors are able to observe attributes of an instance not implied by the class(es) they can provide.

Proposition 2 (Information Loss): Class-based conceptual models result in information loss when the class that a contributor uses to record an instance does not imply some attributes of the instance observed by the contributor.

3.3 Impact of Conceptual Modeling on Dataset Completeness

Whereas information loss deals with the representation of attributes of things, dataset completeness addresses the issue of whether any information about a thing is captured at all. For example, if an online contributor attempts to provide some information about an instance (e.g., product, planet, animal), but the IS rejects the entire attempt resulting in failure to capture any information about the instance, dataset completeness is undermined. Dataset completeness is of critical concern to organizations. Fan and Geerts (2012) warn, "not only attribute values but also tuples are often missing from our databases" (pp. 93-94).

Informing the approach to dataset completeness is the perspective taken by Wand and Wang (1996) who argued that "completeness is the ability of an information system to represent every meaningful state of the represented real world system" (p. 93). Although their analysis is premised on IQ that reflects "the intended use of information" (p. 87), it suggests that dataset completeness maybe undermined if an IS is incapable of representing every potentially relevant state of the world.

This thesis argues class-based modeling negatively impacts dataset completeness due to the requirement to comply with the constraints specified in class-based conceptual models. For example, an instance will be rejected by an IS if a class a contributor wishes to use to report the instance is not specified in the conceptual model. Similarly, if, when

reporting an instance of a class, some attributes do not match those defined by the IS, the entire instance may be rejected. This places unnecessary limitations on providing information especially in domains such as UGC where completely specifying the relevant classes in advance is unrealistic. Furthermore, a mismatch between models of a contributor and those defined in the IS may dissuade data contributors from reporting information. For example, users may be apprehensive of submitting potentially incorrect data (e.g., an instance of an animal for which no specific class is found), or even be frustrated by the gulf between his or her own model and that reflected in the IS and thus avoid using the system.

Proposition 3 (Dataset Completeness): Class-based conceptual models undermine dataset completeness (resulting in fewer instances stored) when the classes defined in an information system do not match those familiar to the information contributor.

3.4 Chapter Conclusion

This chapter provided a theoretical foundation for crowd IQ and conceptual modeling. Specifically, it leveraged theories in philosophy and psychology to derive propositions about the impact of conceptual modeling on important IQ dimensions of accuracy and completeness (including information loss and dataset completeness). These provide the basis for testable propositions that this thesis evaluates in laboratory and field settings in subsequent chapters.

The next chapter presents three laboratory experiments that examine the impact of class-based conceptual models on accuracy and information loss in the context of UGC.

Chapter 7 presents a field experiment in the context of citizen science in biology to test the relationship between conceptual modeling approaches and dataset completeness.

4 Impact of Conceptual Modeling on Accuracy and Information Loss

4.1 Introduction

As outlined in Chapter 1, UGC is rapidly becoming a valuable organizational resource. In many domains – including business, science, health and governance – UGC is seen as a way to expand the scope of information available to support decision making and analysis. To make effective use of UGC, understanding and improving crowd IQ is critical. Traditional IQ research focuses on corporate databases, and views users as data consumers. However, as users with varying levels of knowledge or expertise increasingly contribute information in an open online setting, current conceptualizations of IQ break down.

The previous chapters introduced the concept of crowd information quality (crowd IQ), and proposed the impact of traditional class-based modeling approaches on crowd IQ. In particular, I argued that the traditional practice of modeling information requirements in terms of a fixed structure of classes, such as an Entity-Relationship diagram or relational database tables, unnecessarily restricts the level of IQ that can be achieved in user-generated datasets. To evaluate these propositions regarding accuracy and completeness (information loss) in UGC, I conducted three laboratory experiments in the context of a citizen science project in the natural history domain. Citizen science epitomizes the concept of UGC (Hamel et al. 2009; Hochachka et al. 2012; Kim et al. 2011; Wiggins et al. 2011). Citizen science is a type of crowdsourcing in which scientists

enlist ordinary people to generate data to be used in scientific research (Louv et al. 2012; Silvertown 2009). Citizen science promises to reduce information acquisition costs and facilitate discoveries (see, for example, Hand 2010).

Citizen science in biology is a convenient ground for research in IQ: it has established standards for information quality (e.g., biological nomenclature) and a well-defined cohort of data consumers (scientists). This makes it easier to evaluate the impact of modeling approaches on real decision making. Further, citizen science has an immutable requirement for high-quality data - an important requisite for valid research. Citizen science is a voluntary endeavor and the challenge is to induce data of acceptable quality while keeping participation open to broad audiences (Louv et al. 2012).

Within the broader context of citizen science, biology has a well-established conceptual schema. Specifically, *species* is considered the focal classification level into which instances in this domain are commonly organized. Species are units of research, international protection and conservation (Mayden 2002). Major citizen science projects (e.g., eBird.org collecting millions of bird sightings) implement prevailing modeling approaches (e.g., Entity-Relationship) and collect observations of instances as biological species (Parsons et al. 2011; Wiggins et al. 2013).

Major science projects, such as eBird (see Table 1) focus on species identification and advocate Entity-Relationship Diagrams as “best practice” for modeling citizen science domains (Wiggins et al. 2013). Therefore, evaluating the impact of class-based models on the quality of contributions in these projects is of great practical importance.

Table 1. Major citizen science projects that harness UGC

Project	Scope	Collection focus*	No of records **
eBird www.ebird.org	Birds, globally	Species-level	Over 100 million
The Atlas of Living Australia http://www.ala.org.au/	All taxa, Australia	Species-level	Over 35 million
iSpot http://www.ispotnature.org/	All taxa, globally (UK primarily)	Species-level	Over 250,000
South Asia Birds http://www.worldbirds.org/	Birds, India primarily	Species-level	Over 50,000
Treezilla http://www.treezilla.org/	Trees, UK	Species-level	48,000

*Projects may allow other levels, but species is the principal level at which data collection is expected. **As of May. 2014; records come from various sources (e.g., citizens, experts, and existing collections).

4.2 Experiment 1

4.2.1 Impact of Conceptual Modeling on Accuracy in a Free-form Data Collection

First, I investigate the impact of conceptual modeling on accuracy and information loss in a free-form data reporting task. While users typically select from predefined classes, a free-form task makes it possible to investigate the impact of modeling on IQ in the absence of potential confounds arising from guiding participants to particular classes (e.g., priming, cuing effects). The unprompted setting enables exploration of the kinds of classes and attributes contributors naturally choose when describing familiar and unfamiliar phenomena (in Experiments 2 and 3 in this chapter, I guide participants to predefined classes).

Information systems supporting many natural history citizen science projects are class-based and involve positive identification (i.e., classification) of *genera* or *species* (Parsons et al. 2011; Silvertown 2010), as this information is demonstrably useful for scientific research (Bonter and Cooper 2012). Therefore, data collection involves

classifying observations at the species-genus level and contributors are presented with options based on this conceptual model (see Table 1).

However, citizen scientists generally are not biology experts.⁴ In general, I expect individuals with low expertise to have limited skill in identifying species, and to be only able to correctly identify relatively few, widely known (familiar) species. Requiring contributors to classify observations at the species-genus level may lead to guessing and, thereby, result in inaccurate data. As an alternative, the *basic level* is widely accepted in cognitive psychology as the generally preferred classification level for non-experts (Rosch et al. 1976). In biology, the basic level is an intermediate taxonomic level (e.g., “bird” is a level higher than “American Robin”, and lower than “animal”). Jolicoeur et al. (1984) suggest the basic level is typically the first class people think about when they encounter an instance. Children appear to learn basic level classes ahead of other classes, and people use them most frequently in daily speech (Cruse 1977; Murphy and Wisniewski 1989). Experimental studies have shown that people are generally able to

⁴ Defining expertise is not straightforward and not necessarily based on formal credentials. An individual may be recognized as an expert in one domain, but not in another, similar one. Expertise is also likely to exist along a continuum rather than as a binary condition (Collins and Evans 2007). This thesis considers expertise as the *level of contributor domain knowledge relative to an intended use of information as determined by project sponsors*. In the case of natural history citizen science, this can be operationalized as species identification skill.

classify objects more quickly (e.g. Murphy 1982) and more accurately (e.g. Rosch et al. 1976) at the basic level than at subordinate or superordinate levels.

The contrast between basic and species-genus levels clearly illustrates the potential mismatch between the classification structure of a contributor and the one defined in an IS, resulting in a potential deterioration of data quality (Proposition 1, Chapter 3).⁵ As the expected preferred level for non-experts is the basic level, I therefore expect that, in an unprompted setting (i.e., participants do not choose from a pre-determined set of classes), non-experts will *classify more often and more accurately* at the basic level than at the species-genus level. This leads to the following hypothesis:

H-1.1 (Information Accuracy). In a free-form data entry task, contributors will classify instances with higher accuracy (fewer errors) at the basic level than at the species-genus level, when classes at the species-genus level are unfamiliar to the contributors.

⁵ Proposition 1 (Classification Accuracy) states that class-based conceptual models result in lower information accuracy (more classification errors) when the classes defined in an information system do not match those familiar to the information contributor

4.2.2 Impact of Conceptual Modeling on Information Loss in a Free-form Data Collection

Although basic level classes are expected to increase crowd IQ by producing higher (classification) accuracy from non-expert contributors (by matching classification levels familiar to contributors), the question also arises “to what extent does basic level classification result in information loss?” Following Bunge (1977) and cognitive principles (and consistent with Proposition 2, Chapter 3)⁶, I expect that contributors will tend to report attributes that describe particular instances, rather than attributes associated with a specific class (including a basic level one). For example, when describing a bird (e.g., American Robin, Caspian Tern), I expect non-experts will tend to focus on observable attributes *of the instance*, such as “standing on the ground,” and “orange beak,” as opposed to those associated with its basic level, bird (i.e., “can fly,” “has feathers”). This can be generalized to the claim that a conceptual model based on a particular class level (however useful or intuitive it may be) can preclude (potentially useful) instance-level properties from being recorded, thereby contributing to lower crowd IQ by failing to accommodate the phenomena of interest as perceived by information contributors. This leads to the following hypothesis:

⁶ Proposition 2 (Information Loss) states that class-based conceptual models result in information loss when the class that a contributor uses to record an instance does not imply some attributes of the instance observed by the contributor.

H-1.2 (Information loss). In a free-form data entry task, contributors will describe instances using terms that include attributes subordinate to the level of the class at which they can identify instances.

4.2.3 Experiment 1 Method

To test these hypotheses, I conducted a study with 247 undergraduate business students (141 female, 106 male) in eight experimental sessions at Memorial University of Newfoundland. Participants in each session were shown the same set of stimuli, with the sequence randomized between sessions to mitigate any order effect. Business students were chosen to ensure a low overall level of *biology expertise*, reflecting the intended context where information contributors are non-experts with respect to the intended information uses of project sponsors (in this case, biologists). Low domain expertise was verified using self-reported expertise measures: most participants (83%) either strongly or somewhat disagreed (on a 5-point scale) with the statement that they are “experts” in local wildlife (mean=1.90; s.d.=0.886). Most participants (77%) had never taken any

post-secondary biology courses.⁷ Participants indicated that they spend an average of 10 hours per week outdoors (s.d. = 9.038).⁸ Moreover, the structure of the undergraduate business program did not include formal training in conceptual modeling.

Participation was voluntary and anonymous. Participants were selected from senior business courses and were told the purpose of the study only at the beginning of the session to ensure nobody could prepare in advance and to prevent bias that might arise from attracting students with specific interest in the subject, and vice versa. No incentives (e.g., to encourage correct answers) were provided.

While students are a relatively homogeneous group and unrepresentative of the broader citizen science population, the use of this group as study participants is appropriate. The hypotheses tested are assumed to be universally applicable, as they are derived from fundamental principles of human cognition. The participants were selected with low biology expertise because those with little domain knowledge may be most

⁷ While the demographic data indicate an overall low level of biology expertise among participants, 47 participants reported they had taken more than one course in biology and 12 participants strongly or somewhat agreed with the statement that they were “experts” in local wildlife. To justify using these participants together with the rest of the sample in the test of accuracy (H-1.1), I compared the number of correct responses at: (1) species/genus and (2) basic levels, between non-experts and these potential experts. Welch’s *t*-test showed no significant difference between the groups (*p*-values of 0.11 and 0.81); therefore, I used the full sample in further analysis.

⁸ Finally, the low proportion of species-level responses obtained in Experiment 1 (discussed below) is further evidence of low expertise.

disenfranchised in UGC designed based on class-based conceptual models. Furthermore, students can be good predictors of where the rest of the society is moving vis-à-vis information technology adoption (Gallagher et al. 2001).

4.2.3.1 Materials

The stimuli were 24 full-color images of plants and animals (see Appendix 1) native to Newfoundland and Labrador. The plants and animals were selected by an ecology professor well-versed in flora and fauna of the region. Species were chosen to include some organisms believed to be familiar and some believed to be unfamiliar to people living in the area. In each image, the organism of interest was in focus and occupied most of the image area.

Participants were randomly assigned into one of two study conditions. Those in the first condition (Categories and Attributes; 122 participants) were given a printed form with two columns - one asking participants to *name* the object on the image (using one or more words) and the second asking them to *list features that best describe the object* on the image. In the second condition (Attributes only; 125 participants), there was only one column asking participants to *list features that best describe the object*.

4.2.3.2 Procedure

Images were displayed to participants in a random sequence on a large screen. Each image was shown for 50 seconds. This time was deemed reasonable, as observers often have only short encounters with fauna in the wild, and in a pre-test it was

determined sufficient to elicit several attributes and classes. The transition between images was a blank screen shown for one second, accompanied by a beep.

4.2.3.3 Data Entry

I transcribed the responses to ensure consistency. I recorded verbatim the categories and attributes provided by participants, following practices used in similar studies (Jones and Rosenberg 1974; Lambert et al. 2009). When faced with illegible handwriting I attempted to decipher handwriting but avoided making interpretations and skipped unreadable entries. Obvious spelling errors were corrected (e.g., *coyotai* was coded as *coyote*); redundant words (e.g., *its antlers look heavy* was coded as *heavy antlers*) and symbols (e.g., brackets, tilde) that did not carry additional meaning were removed. Complex attributes were broken down into individual components (e.g., “long yellow beak” was coded as “long beak” and “yellow beak”), based on considerations suggested by Rosenberg and Jones (1972). Following psychology research (e.g., Tanaka and Taylor 1991), attributes for the same species with clearly similar meanings were grouped together (e.g., “horns” and “antlers”).

4.2.3.4 Coding

Categories were coded as either “basic level,” “species-genus level,” or “other” and attributes as either “basic level,” “superordinate to basic,” “subordinate to basic,” or “other.” The species-genus level was determined based on biological convention, while the basic level was adopted from prior studies in cognitive psychology (Klibanoff and Waxman 2000; Lassaline et al. 1992; Mervis and Crisafi 1982; Michael et al. 2008;

Murphy 1982; Rhemtulla and Hall 2009; Rosch 1974; Tanaka and Taylor 1991). All categorical responses at other biological levels (e.g., subordinate) were coded as “other”. A thorough survey of cognitive literature failed to reveal an agreed-upon basic-level for 6 out of the 24 species used (lung lichen, Old Man’s beard, coyote, chipmunk, moose, and caribou), so these were excluded from further analysis. The final data set contained 3,737 categories and 7,330 attributes.

For internal consistency, I coded all the data. To assess coding accuracy, another person independently recoded category responses, resulting in 94.8% agreement with the original coding (Cohen’s Kappa = 0.913). This agreement is considered “almost perfect” (Landis and Koch 1977). The third individual independently recoded the attributes, with 76.3% agreement⁹ with the original coding.

⁹ Cohen’s Kappa for attributes was 0.209, which is borderline “fair agreement” (Landis and Koch 1977). The decrease in Kappa is due to the high prevalence of subordinate attributes which, according to both coders, accounted for at least 74% of all attributes (*prevalence index* = 0.66, which is considered high, see Sim and Wright 2005). Coders agreed on what to code as “subordinate” 86.6% of the time, but the pervasiveness of subordinate attributes influences the Kappa statistic as an indicator of chance agreement (Sim and Wright 2005). In cases of high prevalence, raw agreement and prevalence index tend to be more informative than Kappa values (Sim and Wright 2005). All indicators are consistent with the hypothesis H-1.2 that predicts more subordinate attributes.

4.2.4 Experiment 1 Results

4.2.4.1 Information Accuracy: Free-form Data Entry (H-1.1)

To assess accuracy, I focused on the “Categories and Attributes” study condition, in which 122 participants were explicitly asked to classify observed stimuli. Participants provided a total of 3,737 categories (on average 1.28 per image per participant). I analyzed data for each image separately. The categories for each species were grouped into basic and combined species-genus levels (categories at other levels were not relevant this analysis). The basic level (e.g., *bird*) was expected to be preferred by participants, while species (e.g., American Robin, *Turdus migratorius*) and genus (e.g., “true thrush,” *Turdus*) levels are useful to data consumers (e.g., biologists) and are the levels at which many citizen science projects expect contributors to report sightings.

As expected, basic-level categories were most frequent. To compare the frequency of basic and species-genus level responses, the Chi-square goodness of fit statistic was used. The observed frequencies of basic and species-genus labels were compared with the null model assuming equal proportions of basic and species-genus level categories (aggregating species and genus categories into one group increased the test’s conservativeness). For example, when observing Common Tern, participants provided 107 basic level (e.g., *bird*) and 3 species-genus level responses. The expected frequency for each group is 55 ($\chi^2=98.33$, d.f.=1, $p < 0.001$). This shows a strong tendency to report basic-level categories, consistent with prior research in cognitive psychology.

Table 2. Chi-square (χ^2) goodness-of-fit for the number of basic vs. species-genus level categories

Species	Basic and species-genus	Basic	Species-genus	Ratio of basic to species-genus	χ^2	p-value
American Robin	164	86	78	1.10	0.39	0.532
Atlantic salmon	125	100	25	4.00	45.00	0.000
Blue Jay	168	69	99	0.70	5.36	0.021
Blue Winged Teal	149	144	5	28.80	129.6	0.000
Bog Labrador tea	112	108	4	27.00	96.57	0.000
Calypso orchid	104	92	12	7.67	61.54	0.000
Caspian Tern	113	111	2	55.50	105.1	0.000
Common Tern	110	107	3	35.67	98.33	0.000
False morel	34	34	0	N/A	34.00	0.000
Fireweed	120	94	26	3.62	38.53	0.000
Greater	109	108	1	108.00	105.0	0.000
Indian pipe	96	89	7	12.71	70.04	0.000
Killer whale	142	54	88	0.61	8.14	0.004
Mallard Duck	153	133	20	6.65	83.46	0.000
Red fox	124	110	14	7.86	74.32	0.000
Red squirrel	123	105	18	5.83	61.54	0.000
Sheep laurel	105	103	2	51.50	97.15	0.000
Spotted Sandpiper	114	112	2	56.00	106.1	0.000

Table 2 summarizes the results. In 15 of 18 images, there was a significant ($p < 0.001$) preference for basic-level categories.¹⁰ Only in the case of American Robin, killer whale and Blue Jay did basic-level classification not dominate. In the case of killer whale and Blue Jay, participants favored the species, rather than the basic, level (bird or whale).

¹⁰ Allowing for multiple comparisons (18 in this case), a Bonferroni correction can be made to calculate a more conservative p-value ($.05/18=.0028$). Note that the results are robust to this adjustment, as the significant results favoring basic-level categories remain significant.

This can be explained by the familiarity with these animals among participants. The prevalence of basic-level category responses across most of the stimuli is further evidence of the low level of domain expertise in the sample.

Table 3. Fisher's exact test of independence in Categories and Attributes condition

Species	Correct basic	Incorrect basic	Correct species-genus	Incorrect species-genus	Fisher's exact (<i>p</i> -value)
American Robin	86	0	74	4	0.049
Atlantic salmon	100	0	0	24	0.000
Blue Jay	69	0	98	1	1.000
Blue Winged Teal	143	1	0	5	0.000
Bog Labrador tea	108	0	0	4	0.000
Calypso orchid	91	1	0	12	0.000
Caspian Tern	111	0	0	2	0.000
Common Tern	107	0	0	3	0.000
False morel	22	12	0	0	N/A
Fireweed	94	0	1	25	0.000
Greater Yellowlegs	107	1	0	1	0.018
Indian pipe	88	1	0	7	0.000
Killer whale	48	6	86	2	0.054
Mallard Duck	133	0	15	5	0.000
Red fox	104	6	10	4	0.015
Red squirrel	100	5	1	17	0.000
Sheep laurel	103	0	0	2	0.000
Spotted Sandpiper	112	0	0	2	0.000

To test *accuracy* (H-1.1), I assigned a binary variable for each response indicating whether it was correct for the stimulus it described. For example, in descriptions of *Common Tern*, all labels *bird* were coded as correct (at the basic level); *Common Tern* was coded as correct at the species-genus level, while *Arctic Tern*, *Kittiwake*, and *Osprey* were coded as incorrect. I performed Fisher's exact test of independence to determine if

information accuracy was contingent on level of classification. As show in Table 2, for half of the images very few species-genus level categories were provided.¹¹

The results are significant (using a threshold of $p=0.05$) for 15 out of 17 species (excluding False morel, for which a p value could not be calculated due to a complete absence of species-genus level responses, while 22 participants correctly provided its basic level, *mushroom*), indicating a strong relationship between level of classification and accuracy (see Table 3).¹² In all significant cases, the number of correct basic level responses was higher than the number of correct species-genus level responses. The cases for which accuracy was not significantly higher for basic level categories (i.e., Blue Jay and killer whale) involved familiar or commonly known species that non-experts may see often, either in nature or in the media. It is reasonable to postulate that high prior exposure to these species resulted in high accuracy at the species level, and these two species accounted for a high proportion of all correct species-genus level responses. Notwithstanding these charismatic cases, the remainder of the data demonstrates that, as

¹¹ Fisher's exact test was chosen over Chi-square due to low frequencies in species or genus cells. Unlike Chi-square, Fisher's exact test provides exact hypergeometric probability (expressed as a p -value) of observing this particular arrangement of the data. Despite criticisms of being unnecessarily conservative, it remains a popular method to detect contingency in categorical data and is preferred in data with low expected cell values (Agresti 1992).

¹² Allowing for multiple comparisons (17 in this case), a Bonferroni correction can be made to the p -value ($.05/17=.0029$). The results are robust, favoring basic-level categories in 12 of the 17 cases.

the level of classification changes from basic to species-genus, accuracy declines. Overall, the results provide strong support for H-1.1.

4.2.4.2 Information Loss (H-1.2)

I measured information loss in terms of the number of attributes reported by participants that could not be inferred from the classes provided by those participants for an image. The results from the accuracy test above demonstrate the dominant performance of basic level categories over species-genus level categories. This finding is critical in testing the degree of information loss, as the question can now be asked “to what extent do participants employ basic-level attributes (e.g., *can fly*, *has feathers* for *bird*) versus lower-level attributes (e.g., *red breast*) when they are not required to classify observations?” The greater the number of sub-basic level attributes reported, the greater the degree of potential information loss if the basic level is the one at which information is collected and stored.

To investigate information loss, all attributes (7,330) in the Attributes-only condition for the 18 plants and animals with an agreed-on basic level category were classified into: sub-basic, basic (and superordinate), or other, resulting in 6,429 sub-basic, 824 basic, and 77 other attributes. Table 4 illustrates the sub-basic, basic and other attributes provided for one of the organisms in the study (American robin).

Table 4. Sample of basic, sub-basic and other attributes provided for American robin in the Attributes-only condition

Frequency count	Basic	Sub-basic	Other
85		red breast	
31		small	
26		yellow beak	
22	has feathers		
20		black	
15		black head	
14		small beak	
12		brown	
9		pointy beak	
9		black back	
8	can fly		
...
1			never seen before

I tested for differences using the Chi-square goodness of fit test, where the observed frequencies of sub-basic and basic level attributes were compared with expected frequencies (assuming equal probabilities of obtaining basic and sub-basic attributes). In contrast with the prevalence of basic level categorization, there were 9.38 times more sub-basic than basic level attributes, with an average *p*-value approaching *zero*. Table 5 summarizes the results across the 18 species used in this analysis. The data strongly support H-1.2 and indicate that, despite the salience of a particular classification level, the basic-level does not capture all information available to and easily reported by contributors.

Table 5. Number of sub-basic, basic, super-basic and other attributes in Attributes-only condition

Species	Total	Sub-basic	Basic	Sub-basic to basic ratio	Super-basic	Other*	χ^2 p-value (basic and super vs. sub-basic)
American Robin	400	362	35	10.3	1	2	0.000
Atlantic salmon	337	273	45	6.1	4	15	0.000
Blue Jay	453	397	51	7.8	1	4	0.000
Blue Winged	439	350	76	4.6	2	11	0.000
Bog Labrador tea	274	266	3	88.7	2	3	0.000
Calypso orchid	364	358	3	119.3	0	3	0.000
Caspian Tern	511	460	47	9.8	1	3	0.000
Common Tern	479	435	41	10.6	0	3	0.000
False morel	248	238	9	26.4	0	1	0.000
Fireweed	312	302	3	100.7	0	7	0.000
Greater	534	486	39	12.5	4	5	0.000
Indian pipe	351	342	6	57.0	0	3	0.000
Killer whale	388	325	54	6.0	0	9	0.000
Mallard Duck	497	421	74	5.7	0	2	0.000
Red fox	476	340	46	7.4	88	2	0.000
Red squirrel	503	362	105	3.4	35	1	0.000
Sheep laurel	326	319	4	79.8	0	3	0.000
Spotted	438	393	44	8.9	1	0	0.000

*Some attributes provided could not be associated with biological classes of organisms. For example, some participants used adjectives such as “beautiful” and “standing on rock” to describe organisms.

4.3 Experiment 2

In Experiment 1, the classes that would be of interest to project sponsors did not in most cases match contributor classifications of phenomena in the domain. However, the experimental task did not direct participants to a particular level of classification. In practice, data collection (whether for UGC or traditional applications) typically involves populating pre-existing class structures. Experiment 1 demonstrates that class-based models can impair accuracy and result in information loss, but does not provide direct evidence of the impact of a predefined schema (i.e., when classes are predefined in

advance and contributors are asked to select among these classes) on accuracy. Hence, I conducted a second experiment to assess whether the findings from Experiment 1 (free-form) change when a predefined class-based schema is imposed.

In Experiment 2, participants classify each stimulus by selecting one option from pre-specified options. Based on the results of Experiment 1, the classification choices (levels) available to participants were manipulated. The first condition simulated a class-based model at a single (species) level, typical of existing projects (i.e., select a species from a set of potential species). The second condition simulated a hierarchical class-based model (e.g., species options, as well as superordinate and subordinate classes). In particular, there were correct classes at different levels (e.g., superordinate to basic, basic, subordinate to basic, species). Importantly, each set of classes in this condition included the most frequent (and always correct) response from Experiment 1 (e.g., bird, fish). It also included multiple incorrect options (at different levels) to make the task more realistic (the number of incorrect options varied slightly for different organisms). For example, the options for Common Tern (*Sterna hirundo*) were: animal (correct, superordinate), bird (correct, basic), Common Tern (correct, species-level), Iceland Gull (incorrect, species-level), loon (incorrect, subordinate), shorebird (incorrect, subordinate), tern (correct, subordinate), warm-blooded organism (correct, superordinate), waterfowl

(incorrect, subordinate).¹³ In addition, each condition included “I don’t know” and “Other” (with space for an alternate response) options to allow participants to either avoid classifying (typical to volitional IS use) or respond using classes that were not among the predefined choices.

Experiment 1 showed that non-experts favor basic level classes. Therefore participants are expected to *classify more often and more accurately* at the basic level, leading to higher accuracy in the multi-level condition, where the basic-level option is explicitly provided as one of the options. Consistent with Proposition 1, this leads to the following hypothesis:

H-2 (Information Accuracy). In a constrained (class-based) data entry task, contributors will classify instances with fewer errors in a multi-level (super-, basic- and sub-basic) model than in a single-level (species-genus) model, when the classes in the single-level model are unfamiliar to the contributors.

4.3.1 Experiment 2 Method

Seventy seven undergraduate students (24 female, 53 male) participated in the study. Almost all (94.8%) strongly or somewhat disagreed (on a 5-point Likert scale) with

¹³ A complete listing of options provided to participants for all species used is provided in Appendix 2.

the statement that they are “experts” in local wildlife, and most (68.8%) had never taken a post-secondary course in biology.

4.3.1.1 Materials and Procedure

The materials used were a subset of those in Experiment 1.¹⁴ The procedure for presenting the images was the same as in Experiment 1. Participants were randomly assigned into one of two conditions. In the single-level condition (38 participants), participants chose from a list of possible species-level responses; in the multi-level condition (39 participants), participants chose from options that included the basic level and levels above and below the basic (including species). Nothing in the study materials suggested that the responses were required at a particular (i.e., specific or more general) level.

In the single-level condition, of the nine species provided as options, only one was correct. The eight others were selected as plausible options based on similarity in appearance and/or habitat, and their occurrence in the same geographic region. In the multi-level condition, the options were selected based on Experiment 1 to increase congruence with non-expert classifications. There was the same number of

¹⁴ Experiment 2 excluded a number of images used in Experiment 1 (see Appendix 1) – those for which there is no agreed-on basic-level category (e.g., lung lichen, Old Man’s beard), and those familiar species that participants were able to identify correctly in Experiment 1 (i.e., American Robin, Blue Jay, killer whale).

correct/incorrect options across all ten images. The full list of options presented to participants is listed in Appendix 2. The options were printed on paper with each set of options on its own page. In both conditions the order of options was randomized for each participant, and participants were asked to select one option (the options were not grouped in any way and the classification level was not indicated). In addition to facilitating comparison between groups, the options in the single-level condition were mutually exclusive, while in the multi-level condition, lower level options implied higher level ones (e.g., American Robin implied bird) and options at the same level were mutually exclusive.

4.3.2 Experiment 2 Results

In assessing accuracy, I compared the answers given by participants in the single-level and multi-level conditions. The responses from the predefined list of 9 options and the responses written in “Other” field were combined. The “I don’t know” responses were excluded from the count – making the test more conservative (there were 108 “I don’t know” responses in the single-level condition and only 15 in the multi-level condition). In total, 271 responses in the single-level condition were compared with 375 responses in the multi-level condition. Each response was coded as “correct” or “incorrect” based on biological convention (e.g., the answer *bird* was accurate for *Common Tern*, but *seagull* was inaccurate).

Table 6. Comparison of accuracy in Experiment 2: single (E2SL) vs. multi-level conditions (E2ML)

Species	E2SL			E2ML			E2ML vs. E2SL		
	Correct	Incorrect	% Correct	Correct	Incorrect	% Correct	% Diff.	χ^2	p-value
Atlantic salmon	10	23	30.3	32	7	82.1	51.7	19.694	0.000
Blue Winged	6	27	18.2	32	7	82.1	63.9	29.257	0.000
Calypso orchid	7	17	29.2	29	8	78.4	49.2	14.576	0.000
Caspian Tern	4	20	16.7	23	15	60.5	43.9	11.510	0.001
Common Tern	5	22	18.5	22	17	56.4	37.9	9.476	0.002
False morel	0	24	0.0	30	4	88.2	88.2	43.866	0.000
Fireweed	7	17	29.2	29	10	74.4	45.2	12.390	0.000
Indian pipe	4	21	16.0	16	20	44.4	28.4	5.417	0.020
Mallard Duck	26	11	70.3	36	3	92.3	22.0	6.136	0.013
Sheep laurel	4	16	20.0	28	7	80.0	60.0	18.832	0.000
AVERAGE			26.9		73.9		47.0		

As expected, accuracy in the multi-level condition was significantly greater than in the single-level condition (73.9% versus 26.9%, $p=0.000$, $\chi^2=139.56$, 1 d.f.). This was largely due to the prevalence of correct responses at the basic level in the multi-level condition: there were more basic-level (148 or 39.5%) than species-level (103 responses, 27.5%) responses ($p=0.005$, $\chi^2=8.07$, 1 d.f.). Accuracy of basic-level responses was 99.3% compared with 53.4% for species-level responses. Basic-level responses accounted for 53.1% of correct responses in the multi-level condition (while only 20.2% of correct responses were at the species-level, 7.6% at the subordinate level, and 19.1% at the

superordinate level).¹⁵ To test if the results varied across species, the Chi-square goodness of fit statistic was computed for each pair (Table 6). In all cases, accuracy in the multi-level condition was significantly greater than in the single-level condition.¹⁶ These results strongly support H-2 (and are consistent with H-1.1).

4.4 Experiment 3

Experiments 1 and 2 demonstrate that accuracy declines if the classes specified in a conceptual model do not match the classes contributors are able to provide competently. Experiment 3 sought to rule out possible alternative explanations for the finding in Experiments 1 and 2. First, it was necessary to ensure that participants in the species-level condition were not drawn to incorrect options merely due to greater familiarity with these options than with the correct one. Therefore, I examined the results of Experiment 2 and removed and replaced all incorrect classes that received a larger than average number of responses (a possible indicator of participant familiarity with these options). For example,

¹⁵ Greater accuracy in the multi-level condition was not merely a function of the number of correct options available in the single-level condition (one correct response) versus the multi-level condition (several correct responses). While most options available were at levels other than basic, participants consistently favored the correct basic option and avoided other levels (including incorrect basic, species, superordinate). A detailed analysis of the responses is provided in Appendix 3.

¹⁶ Allowing for multiple comparisons (10 in this case), a Bonferroni correction can be made to calculate a more conservative p-value ($.05/10=.005$). Note that the results are robust to this adjustment, with 8 of 10 cases remaining significant.

Jelly leaf fungus was removed as an option for False morel because it was incorrectly chosen 13 times in Experiment 2, whereas the next most frequent incorrect response was selected 5 times. All frequent incorrect responses were replaced with new classes deemed by the ecology professor (who selected options in Experiment 1) to be unfamiliar to non-experts.

Second, to ensure that the results in Experiment 2 were not influenced by omitting the species from Experiment 1 that were known to participants, Experiment 3 added the species from Experiment 1 that were removed in Experiment 2 (i.e., American Robin, killer whale and Blue Jay). Including these created a familiar (or “schema-congruent”) set of stimuli, based on the finding from Experiment 1 that participants were able to identify these organisms at the species level and on research on basic-level categorization showing participants prefer more specific classification when they are experts in a domain (Tanaka and Taylor 1991). This “schema-congruent” set could be compared with an unfamiliar (“schema-incongruent”) group – the 10 classes from Experiment 2 for which accuracy was greater in the multi-level condition. Consistent with Proposition 1 and H-2, this leads to the following hypothesis:

H-3.1 (Information Accuracy). In a constrained (class-based) data entry task, contributors will classify instances with fewer errors in a multi-level (super-, basic- and sub-basic) model than in a single-level (species-genus) model, when classes in the single-level model are unfamiliar to the contributors.

Finally, to further evaluate the claim that requiring non experts to conform to a predetermined class-based schema has negative consequences on IQ, I compare

classification accuracy in free-form vs. constrained data entry tasks. While constrained data entry provides participants with cues and may help in recalling applicable classifications, it may also bias participants to choices they might not otherwise make, leading to wrong classification decisions (Parsons et al. 2011). For example, whereas non-experts can provide accurate responses in a free-form data task (as seen in Experiment 1 where the overall accuracy of categories provided was 86.7%), the presence of different options may influence data contributors to select incorrect classes. Consistent with Proposition 1, this leads to the following hypothesis:

H-3.2 (Information Accuracy). In a free-form data entry task, contributors will classify instances with fewer errors than in a constrained (class-based) data entry task, whether the latter uses single-level or multi-level classification, when classes at the species-genus level are unfamiliar to the contributors.

4.4.1 Experiment 3 Method

Sixty six undergraduate business students (36 female, 30 male) participated, drawn from the same population of biology non-experts as in Experiments 1 and 2. Almost all participants (89.4%) strongly or somewhat disagreed (on a 5-point Likert scale) with the statement that they were “experts” in local wildlife, and most (83.3%) had never taken a post-secondary course in biology.

4.4.1.1 Materials and Procedure

The materials used were the same as in Experiment 2, with the addition of the three familiar species used in Experiment 1. The procedure for presenting the images was

the same as in Experiments 1 and 2. Participants were randomly assigned into one of three conditions. In condition 1 (23 participants), participants chose one option from a list of possible species-level responses. In condition 2 (21 participants), participants chose one option from classes at the basic level and at levels above and below the basic (including species). In both conditions, “I don’t know” and “Other” (with space for an alternate response) options were included to allow participants to either avoid classifying or respond using classes that were not included in the predefined lists. In condition 3 (22 participants), participants were presented with an empty sheet and asked to name the object using one category or write "I don't know".

In the single-level condition, of the nine species provided as options, only one was correct. The eight others were selected as plausible alternatives based on similarity in appearance and/or habitat, and their occurrence in the same geographic region. In the multi-level condition, there were four correct (including the most frequent correct responses from Experiment 1, such as fish, bird, and mushroom) and 5 incorrect options for each species.¹⁷ The options were printed on paper, with each set of options on its own page. In both conditions, the order of options was randomized for each participant and participants were asked to select one option for each stimulus.

¹⁷ Appendix 3 provides detailed analysis showing that the results are not compromised by the potential bias of different numbers of correct responses in the single-level and multi-level conditions.

4.4.2 Experiment 3 Results

4.4.2.1 Impact of Schema on Accuracy: Single vs. Multiple Level Class-Based Model (H-3.1)

In assessing accuracy, the same procedure used to test H-2 was followed. The “I don’t know” responses were excluded from the count, thereby making the test conservative (there were 86 “I don’t know” responses in the single-level condition and 19 in the multi-level condition). In total, 213 responses in the single-level condition were compared with 254 responses in the multi-level condition. Each response was coded as “correct” or “incorrect” based on biological convention.

As expected, accuracy in the multi-level condition was significantly greater than in the single-level condition (71.1% versus 49.8%, $\chi^2=23.48$, 1 d.f., $p<0.001$). Accuracy between conditions did not significantly vary for the three familiar species from Experiment 1: 92.1% in the species-only and 91.9% in the multi-level condition with all responses in the single-level condition and most (85.5%) responses in the multi-level condition being at the species level (see Table 7). Participants were comfortable classifying American robin, Blue jay and Killer whale at the species level, suggesting that the species level was congruent with their mental schema for these organisms. Importantly, classification accuracy was not higher for these species in the single-level condition than in the multi-level condition.

Table 7. Accuracy in Single-level (E3SL) and Multi-level condition (E3ML) for "Familiar" species.

Species	E3SL			E3ML		
	Correct	Total	% Accuracy	Correct	Total	% Accuracy
American Robin	16	19	84.2	16	21	76.2
Species-level	16	19	84.2	12	17	70.6
Other levels	0	0	-	4	4	100.0
Blue jay	23	23	100.0	20	20	100.0
Species-level	23	23	100.0	16	16	100.0
Other levels	0	0		4	4	100.0
Killer Whale	19	21	90.5	21	21	100.0
Species-level	19	21	90.5	19	19	100.0
Other levels	0	0	-	2	2	100.0
TOTAL			92.1			91.9
Species-level			92.1			90.6
Other levels*			-	10	10	100.0

* Consisting of 9 basic-level responses (bird, whale) and 1 superordinate (mammal).

For the remaining, unfamiliar group of species, accuracy was greater in the multi-level than in the single-level condition: 65.1% vs. 32.0% (p -value = 0.000, $\chi^2=36.92$, d.f. = 1). In this group, basic-level responses accounted for 63.2% of correct responses in the multi-level condition (while only 20.8% of correct responses were at the species-level). As in Experiment 2, in the multi-level condition for Experiment 3 basic level categorization largely contributed to the greater accuracy for the unfamiliar group of species as compared with the single-level condition: there were more basic-level responses (82 out of 210 responses or 39.0%), compared to the species-level (61 responses, 29.0%); 30 (14.3%) responses were subordinate, 19 (9.0%) responses were superordinate and 19 (9.1%) were "I don't know". Accuracy of basic-level responses was 96.3% compared to 42.6% at the species-level (accuracy at the subordinate and superordinate levels was 13.3% and 84.2% respectively).

These results support H-3.1 and are consistent with H-1.1 (free-form category elicitation) and H-2, providing additional evidence that accuracy is contingent on providing users with classification structures more congruent with preferred user classification models. The lack of difference among the familiar group is consistent with H-3.1 as it suggests that, for these species, most contributors are comfortable classifying at the species-genus level.

4.4.2.2 Impact of Schema on Accuracy: Free-form vs. Class-Based Models (H-3.2)

In assessing accuracy of free-form versus class-based data collection, I followed the procedure used in testing H-3.1, but coded the “I don’t know” responses as incorrect – making the test more conservative when comparing free-form to the multi-level condition (there were 32 “I don’t know” responses in the free-form condition compared with 19 in the multi-level condition). In total 299 responses in the single-level condition were compared with 273 responses in the multi-level condition and 286 responses in the free-form condition.

Overall accuracy in the free-form condition was 77.3% compared to 35.5% in the single-level condition and 66.7% in the multi-level condition (both are significantly lower than the free-form condition based on Fisher’s exact test, $p < 0.05$). I then investigated the differences for each organism separately. As shown in Table 8, in 9 of 13 cases participants in the free-form condition provided a significantly higher percentage of accurate responses compared to those in the single-level condition.

Table 8. Accuracy in Experiment 3, Single-level condition (E3SL), Multi-level condition (E3ML) and Free-form condition (E3FF). * significant at 0.05, ** significant at 0.01 level (using Fisher's exact test).

Species	E3SL		E3ML		E3FF		Δ % Correct	
	Correct / incorrect	% Correct	Correct/ incorrect	% Correct	Correct/ incorrect	% Correct	E3FF vs. E3SL	E3FF vs. E3ML
American Robin	16 / 7	69.6	16 / 5	76.2	20 / 2	90.9	21.3	14.7
Atlantic Salmon	4 / 19	17.4	13 / 8	61.9	17 / 5	77.3	59.9**	15.4
Blue jay	23 / 0	100.0	20 / 1	95.2	20 / 2	90.9	-9.1	-4.3
Blue Winged	11 / 12	47.8	16 / 5	76.2	22 / 0	100.0	52.2**	23.8*
Calypso Orchid	3 / 20	13.0	12 / 9	57.1	17 / 5	77.3	64.2**	20.1
Caspian Tern	1 / 22	4.3	10 / 11	47.6	18 / 4	81.8	77.5**	34.2*
Common Tern	2 / 21	8.7	8 / 13	38.1	12 / 10	54.5	45.8**	16.5
False morel	0 / 23	0.0	14 / 7	66.7	8 / 14	36.4	36.4**	-30.3
Fireweed	7 / 16	30.4	10 / 11	47.6	14 / 8	63.6	33.2*	16.0
Indian Pipe	1 / 22	4.3	6 / 15	28.6	12 / 10	54.5	50.2**	26.0
Killer Whale	19 / 4	82.6	21 / 0	100.0	20 / 2	90.9	8.3	-9.1
Mallard	19 / 4	82.6	19 / 2	90.5	22 / 0	100.0	17.4	9.5
Sheep Laurel	0 / 23	0.0	17 / 4	81.0	19 / 3	86.4	86.4**	5.4
AVERAGE		35.5		66.7		77.3	41.8**	10.6*

Accuracy in multi-level classification was greater than in single-level classification (as shown in H-2.1 and H-3.1 above). In addition, as Table 8 shows, in 2 of 13 cases accuracy in the free-form condition was significantly higher than in the multi-level condition. In part, the increase in accuracy in the free-form condition is due to the greater accuracy when classifying at the basic level (which was close to 100% correct regardless of condition). There were significantly more basic-level responses in the free-form condition than in the multi-level condition; 158 out of 286 times (55.2%) compared to 91 out of 273 times (33.3%) ($p=0.000$, $\chi^2=27.15$, 1 d.f.). There was only one basic-level response (*duck*) in the species-only condition (provided in the "Other" field).

Considering that overall accuracy in the free-form condition was significantly higher than in both of the constrained-choice conditions, the results support H-3.2.

4.5 Chapter Discussion and Conclusion

This chapter evaluates the impact of conceptual modeling on classification accuracy and information loss. Appendix 4 summarizes the findings of the three laboratory experiments. The results demonstrate that accuracy is contingent on the classes used to model a domain. In free-form data collection, except for familiar organisms, the results demonstrate higher accuracy when using basic-level classification. Similarly, in schema-mediated data collection, the results indicate higher accuracy when data collection is organized in terms of classes at multiple levels (including the basic level) as opposed to a single level.

In addition, the comparison between unconstrained and schema-mediated data collection shows that accuracy does not necessarily improve when intuitive and accurate options are provided for users. Indeed, the overall classification accuracy in the free-form condition of Experiment 3 was significantly greater than in either single or multi-level conditions. This is particularly notable, because the most frequent correct options from the free-form task in Experiment 1 (the basic-level categories *bird*, *fish*, *mushroom*) were available as options in the multi-level condition of Experiment 3, making the comparison more conservative. This further indicates the potential IQ implications of using a predefined schema in UGC settings: while predefined classes provide non-expert data contributors with cues that may guide them to correct choices, they may also bias non-experts to wrong classification decisions.

The experiments point to a data quality dilemma in using class-based models to capture UGC. The classes non-experts are comfortable using tend to be general ones. However, for many applications, more specific classes are required. Experiment 1 shows that, when contributors attempt to classify observations at a lower level, accuracy generally declines. Thus, there is the potential for low accuracy in real-world UGC datasets that rely on specialized classification choices. However, the results also show that participants can contribute substantial amounts of information (attributes) beyond what is implied by the high-level classes to which they can assign an observed phenomenon.¹⁸

While the support for H-1.1, H-2, and H-3.1 (i.e., improved accuracy when classes are congruent with contributor views) demonstrates the merits of using more familiar classes (e.g., basic-level categories) in designing information systems to harness UGC, this thesis also examines an alternative to the class-based approach to harnessing collective intelligence. Based on ontology and cognition, I argue that representing instances in terms of classes results in the loss of potentially valuable properties. The test of H-1.2 demonstrated that a significant number of low-level attributes can be generated by non-expert contributors. These attributes cannot be inferred from the classes that can

¹⁸ An interesting question for future research is whether these attributes can be used to infer more specific classes.

be accurately identified by non-experts. Experiments 1, 2 and 3 show that basic-level categories are generally the most frequently provided and typically most accurate of the classification levels, whether in a free-form or schema-mediated data collection tasks. Notwithstanding this, the results also show that modeling using basic level classes can be expected to lead to a significant loss of properties.

Finally, the results provide an empirical evidence of the advantages of the use-agnostic and contributor-focused crowd IQ in UGC settings. Currently, many UGC projects (e.g., various active citizen science initiatives) focus data collection on classifying phenomena using classes that are useful to data consumers. This research suggests that such approaches not only can sometimes lead to data accuracy problems, but can preclude valuable information from being collected (leading to information loss). The results highlight an opportunity to extract additional data from the crowd that is routinely neglected in applications with fixed classification structure. Chapter 8 discusses further implications of these findings.

The next chapter proposes principles for modeling UGC intended to overcome the negative consequences of traditional conceptual modeling on IQ.

5 Principles for Modeling User-generated Content

The increasing reliance of organizations on UGC challenges long-held propositions about conceptual modeling rooted in the assumptions of traditional (e.g., corporate) domains. As demonstrated in the previous chapter, employing traditional class-based conceptual modeling approaches can have negative consequences for crowd IQ. It appears that the potential of UGC is not being fully realized. Motivated by the findings from the three laboratory experiments presented above, this chapter proposes principles underlying an alternative approach to modeling UGC.

5.1 Emergent Approaches to Conceptual Modeling

Recognizing shortcomings of traditional conceptual modeling, several alternative approaches to modeling dynamic, heterogeneous or distributed information have emerged. One approach is to *reduce* the extent and depth of specifications. For example, models may employ only very basic concepts (McGinnes 2011). This concords with agile development which relies on lightweight (“barely good enough”) models that capture semantics minimally necessary for the next design iteration (Ambler 2003). Here one challenge is to convey essential semantics while keeping models simple and lean (Anwar and Parsons 2010).

Whereas lightweight modeling relies on a small number of “core” constructs, an alternative is to use grammars that capture *extended semantics*. Thus, extensions to popular conceptual modeling grammars have been motivated by the need to support dynamic information (Chen 2006; Liu et al. 1994). For example, in dealing with

unpredictability of heterogeneous information, such extensions may employ probabilistic classification models (Ma and Yan 2008). Prior research considered combining abstraction-based constructs with instances (by showing instances that instantiate classes) (Samuel 2012) and icons depicting stylized and typical examples of the abstract constructs (Masri 2009) to improve domain comprehension and understanding by users.

A growing interest is in *domain ontologies* that can “bridge” different systems and users (McGinnes 2011). These ontologies can be constructed by experts or be “outsourced” to the crowd thus purportedly generating more intuitive representations (Braun et al. 2007; Robal et al. 2007). Indeed, such approaches tend to encapsulate diverse user perspectives and are increasingly prolific. Yet even these models may potentially neglect all valid views and thus have a negative impact on IQ. Furthermore, domain ontologies generally require commitment of parties to a predefined (albeit often flexible) conceptual structure (McGinnes 2011).

Another promising approach is putting the onus of modeling on users by allowing them to dynamically change models (Krogstie et al. 2003; Roussopoulos and Karagiannis 2009). This approach may be combined with lightweight modeling in which only a basic model is developed with the expectation that users update the model. This, however, invites unresolved issues of cooperative schema evolution and concurrent access and modification of schemas (Roussopoulos and Karagiannis 2009). It is also unclear if this approach is scalable online, as some users may lack skills and motivation to create and alter models.

The approaches reviewed above presuppose some *a priori* structures and in this sense may have limitations and IQ consequences similar to those of traditional modeling. A promising approach that does not rely on *a priori* structures is to store information in a flexible data model such as the entity–attribute–value (EAV) model. Resource Description Framework (RDF) data model and Datalog logic programming language implement the EAV (Patel-Schneider and Horrocks 2007). The RDF framework supports current approach to the Semantic web by which *things* and concepts on the web can be described using triplets of subject-predicate-object (Heath and Bizer 2011). In Datalog individuals can be declared without a reference to a class. Datalog can be used to declare and store facts about individuals, such as *married (Mary, John)* that describe relationships between individuals Mary and John. While these approaches appear promising, they also have potential limitations. For example, their simplicity potentially comes at the expense of construct overload (Wand and Weber 1993) - whereby the same construct (e.g., object in the RDF triplet) can be used to express different ontological concepts, such as a thing or a class. Empirical evidence suggests ontological deficiencies (i.e., lack of clarity and expressiveness) lead to lower domain understanding (Saghafi and Wand 2014) and negatively impact beliefs about usefulness and ease of use of the grammars (Recker et al. 2011). The applicability of these approaches to modeling UGC is not well understood. Little theoretical understanding exists for how to employ flexible data models to model UGC. A promising approach to model UGC is MIMIC, which advocates principles of flexible representation based on reference theories of psychology (Parsons 1996). MIMIC is based on classical classification theory in cognitive

psychology and assumes the primacy of instances and attributes over classes. The instance independence makes it possible to describe instances using attributes that do not necessarily exist or comply with existing classification structures. Classes can then be formed by abstracting common attributes of instances. While this model was not explicitly tailored to UGC, its propositions regarding instances and attributes are inherently applicable to these settings (a point considered in section 5.3). At the same time, a number of the propositions in this model may not fit well with the nature of UGC settings.

First, MIMIC is based on the classical theory of concepts - defining concepts as bundles of necessary and sufficient attributes (Estes 1996; Murphy 2004; Parsons 1996; Smith and Medin 1981). This may not be problematic in an environment where shared understanding of how to define a class can be reached and maintained. However, this approach appears limiting in UGC settings, as modern psychology research demonstrates that people generally struggle to define classes (concepts) using necessary and sufficient attributes (Murphy 2004; Rosch 1978).

Second, classes in MIMIC are formed by intension (Kimura et al. 1985) - as sets of attributes. This does not permit users to directly provide classes as descriptors of instances. According to modern psychology, in most cases crowd users are unable to generate necessary and sufficient attributes for the classes that they otherwise may easily provide (e.g., a user may easily provide a class *bird*, but struggle to provide enough attributes for definitive identification of instances as members of this class). Indeed, the laboratory experiments in Chapter 4 demonstrate that non-expert crowds can easily and

with high accuracy classify at generic levels (basic-level categories). Allowing users to attach classes to instances directly can exploit the human innate ability to classify (Berlin et al. 1973) and carries a number of other desirable effects (see section 5.3 for more discussion).

Third, MIMIC was originally created to support traditional IS and not all propositions of the model may be germane to UGC settings. For example, MIMIC distinguishes between structural, relational and behavioral attributes and reserves special operations for each (see Parsons 1996). In a UGC setting, it is unrealistic to expect users to understand the differences between these notions and it may be more appropriate to collapse different notions of attributes into one. Similarly, the provision of principles of "good" classification structures do not appear to be applicable to the content users provide in UGC environments as holding crowd users to these principles is challenging (however these principles appear applicable to scientists working with crowd data and constructing classification structures over UGC - a point further developed in Chapter 8).

This thesis shares the ontological and cognitive foundations of MIMIC and builds upon it, but also considers unique challenges and characteristics of UGC. This leads to proposing principles are more closely tailored to the domain of UGC. The next section provides analysis of UGC settings that informs principles of modeling UGC.

5.2 Challenges of Modeling User-generated Content

This section analyzes modeling challenges in UGC settings where traditional conceptual modeling appears to be ill-equipped. Specifically, it focuses on *online citizen science*, in which scientists seek contributions of ordinary people for research purposes

(Louv et al. 2012; Silvertown 2009). As discussed in Chapters 1 and 4, a major aspect of online citizen science is the democratic nature of participation. While projects are developed primarily to serve the needs of scientists (the subject matter experts), the users or contributors (i.e., citizen scientists) are ordinary people, often lacking subject matter expertise and possessing diverse domain views (Coleman et al. 2009). In addition, many projects require only minimal information in order to participate (e.g., to encourage broader participation and/or comply with anonymity requirements of research protocols). As a result, some requirements and domain knowledge may originate from *system owners or sponsors*, but the actual data are provided by diverse and anonymous users. In this environment modeling must embrace the assumption that it may be impossible to reach every relevant and representative stakeholder, making it difficult to determine appropriate and adequate conceptual structures (e.g., classes, relationship types). Similarly, modeling must account for the possibility that some legitimate users are domain non-experts and may not fully understand or be able to comply with the domain views of others. An emerging modeling challenge is having to represent and encourage diversity of user views.

Modeling challenge #1. Represent and encourage diversity of user views.

The scope of many citizen science projects can be extensive and very complex. For example, iSpot.org.uk collects sightings of all natural history in Great Britain. Similarly, Galaxy Zoo images contain a variety of cosmic objects, some unknown to scientists themselves (Lintott et al. 2009). This means no single user is likely to be an expert in the entire application domain. Online citizen science is increasingly used to

answer emerging questions about material and social phenomena. Similarly, scientists may be interested in unique local knowledge or divergent perspectives. As a result, a particular contribution may involve previously unidentified phenomena (instances), creating a challenge to decide how to model the unknown.

Modeling challenge #2. Represent instances of "unknown" classes.

In many projects, the phenomena about which users supply data may be available *only* to the original contributor (or a handful of people). For example, in projects that map biodiversity, the objects of interest (e.g., birds, animals) may be fleeting with an extremely short exposure time. In such cases, it is difficult to exploit redundancy (Franklin et al. 2011; Liu et al. 2012). The focus on representing individual data points does not align well with traditional notions of unified global schema and modeling abstractions, rather than concrete things.

Many citizen science projects explicitly recognize that purposes and uses of the system maybe be undefined at the onset or change over time. For example, the objectives of the Great Sunflower Project (<http://www.greatsunflower.org>) include evolving questions in ecology (e.g., how often do bees pollinate), social sciences (e.g., does participation in citizen science lead to behavioral changes), and computer science and information systems (e.g., how to design systems to increase data quality) (Wiggins et al. 2013). Consequently, the requirement is for modeling to recognize and support undefined and evolving uses of data. Traditionally, modeling assumed intended uses expressed through predefined abstractions. Recognition of evolving uses, however, suggests that

approaches to modeling, be to the extent possible, use agnostic – thus providing more flexibility in repurposing information based on ad hoc needs.

Modeling challenge #3. Encourage unanticipated uses of data.

Unlike many corporate environments, which can be conceptually “frozen” to develop abstract conceptual structures that represent domains, citizen science projects are inherently open: it appears extremely difficult, if not infeasible, to develop appropriate structures that would be congruent with every potential user (stakeholder) in this setting. A conceptual model representing a domain as perceived by some users may marginalize, bias, or exclude possibly valuable conceptualizations of other users. The incongruence between a model of reality embedded in information systems and the one natural for a particular user may preclude the user from effectively engaging and contributing. One consequence of this is low quality (e.g., accuracy, completeness) of information stored in IS. Another consequence is lower engagement (i.e., psychological reaction) with IS that under-represents perspectives of a particular user. On the other hand, freedom from incongruent structures, simplicity and ease of content creation foster greater usage and creativity in usage of IS (Van Kleek et al. 2011).

Modeling challenge #4. Avoid forcing or biasing user views by predefined structures.

Table 9. Modeling challenges in UGC settings

Challenge	Description
MC 1	Represent and encourage diversity of user views
MC 2	Represent instances of the "unknown" classes
MC 3	Encourage unanticipated uses of data
MC 4	Avoid forcing or biasing user views by predefined structures.

In summary, traditional approaches to modeling appear ill-equipped to address the challenges of UGC environments (summarized in Table 9). In the next section I use fundamental theories of philosophy and psychology to propose principles of modeling intended to address the emergent challenges of citizen science and other UGC settings.

5.3 Principles for Modeling User-generated Content

Modeling UGC environments is difficult using traditional abstraction-driven modeling premised on the *a priori* availability of specifications of the kinds of data users might contribute. The analysis of citizen science domains reveals fundamental limitations of the prevailing abstraction-based approaches to domain representation, including the need for consensus among parties involved in modeling and a relatively clear understanding of and agreement on the uses of data.

Abstraction-based conceptual models depict *stylized* (Kaldor 1961, p. 178) - generalized and simplified - representations of actual complex user experiences and beliefs. Psychologically, abstraction is a mental mechanism essential for humans to survive in a diverse and changing world (Harnad 2005; Lakoff 1987; Parsons and Wand 2008). Conceptual modeling grammars based on representation by abstraction assume

that different models elicited from users will be *reasonably similar* making it possible to create a unified view. UGC environments enable new possibilities in which different users are free to maintain their own view of reality, so that capturing individual views becomes critical. Furthermore, focusing users on any one view biases UGC projects to the view of some users and may preclude other views from being represented.

Ontologically, it can be argued that the world is made of unique objects that humans perceive as stimuli (Bunge 1977; Rosch 1978). Humans create abstractions, such as classes, to capture some equivalence among objects for some purpose (Murphy 2004; Smith and Medin 1981). Psychology research contends that prior experience, domain expertise, conceptualization, and ad hoc utility result in different abstractions of the same domain between contributors and for the same contributor over time (McCloskey and Glucksberg 1978; Murphy 2004; Smith 2005). For example, a citizen scientist may create a class of *oiled birds* to refer to distinct objects (birds) that are covered in oil; this class helps the citizen scientist to communicate vital cues about a potential environmental disaster. The same birds seen a few days earlier could have been classified as *beautiful birds* by a group of tourists or *Double-crested Cormorants* by scientists. Modeling using particular "privileged" classes (e.g., species-level, such as *Double-crested cormorants*) promotes some uses, possibly at the expense of others.

In summary, multiple and unique perspectives are part of human experience; it may not be possible or necessary to achieve an agreement among all parties. In UGC settings, user views may not be static and may frequently change. Finally, recognizing the value of information re-use (as implied by the use-agnostic notion of crowd IQ),

modeling in UGC settings needs to be to the extent possible flexible to accommodate evolving, and even unanticipated, uses of data. To achieve these properties, the foundation of UGC modeling should rely on structures that are *invariant across people and do not assume specific uses*. This leads to the formulation of the first principle:

Principle 1. Modeling UGC should be based on user and use-invariant representations.

This principle is a fundamental departure from traditional conceptual modeling driven by abstractions (Mylopoulos 1998). As abstractions naturally vary across people and uses, they do not satisfy the first principle. To derive user and use-invariant structures, this thesis turns to ontology that studies what exists in the world independent of human observers. Philosophy (in particular, ontology) provides a basis for discussing what exists in reality (March and Allen 2012; Wand 1996). Consequently, this thesis adopts a particular ontology (of Mario Bunge) to generate specific statements about reality that are used as the foundation for modeling UGC.

As discussed in Chapter 3, Bunge (1977) postulates that the world consists of “things” (which can also be thought as instances, objects, or entities). This thesis applies the notion of instances to things in the physical, social and mental worlds (Wand et al.

1995).¹⁹ Examples of instances include specific objects that can be sensed in the physical world (e.g., *this chair*, *bird sitting on a tree*, *Barack Obama*) as well any mental objects humans conceive of (e.g., *specific promise*, *rule of algebra*, *Hamlet*, *Anna Karenina*). The fundamental role of instances is supported in psychology, other reference disciplines and in traditional conceptual modeling grammars. According to psychology, instance representation (e.g., spatiotemporal permanence) is a fundamental mental process (Kahneman 1992; Michael et al. 2008; Scholl 2002). People consider individual stimuli (concrete or imaginary) and use abstraction mechanisms to reason (e.g., predict unobserved features) and communicate about them (Falkowski and Feret 1990; Medin and Schaffer 1978; Nosofsky 1986; Rips et al. 2006). People experience a continuous sensory input (e.g., light falling on retina, sound waves) but then eventually transform it into discrete representations (Harnad 1990). Instances become units of attention (Scholl 2002): humans perceive sensory fields (e.g., visual space) to be made of discriminable objects and an undifferentiated perceptual background (Carey 2009; Kahneman 1992). And attention tends to be “allocated to individual objects that are traced through time and

¹⁹ That instances “exist” in physical reality is widely accepted; there is a debate, however, about the extent to which Bunge’s ontology applies to imaginary and social worlds (Allen and March 2012; March and Allen 2012; Wand and Weber 2006; Wyssusek 2006).

space” (Carey 2009, p. 70). Classification typically happens after the existence of an instance is established.²⁰

Instances may also compose to form complex, composite things (Bunge 1977). For example, a computer is made of a central processing unit, a motherboard, random access memory, storage and other components. Composite things may have different attributes of interest than their constitute things, including emergent properties - those that arise as a result of components being put together. Whether a user chooses to represent things as a simple or composite depends on the situation, views and beliefs of the individual user.

Following Bunge, this thesis argues that an instance is an elementary and fundamental construct and, as a consequence, the objective of modeling is to represent instances as fully and faithfully as possible. This leads to the formulation of the second principle:

Principle 2: Instance should be the primary construct in UGC; instances should be represented independent of any other construct.

According to Bunge, every instance is unique in some way and different individuals fail to share some of their properties (see also Proposition 2 in Chapter 3).

²⁰ Note, however, that existing classes may influence what objects in the world are recognized and attended to.

Properties are always attached to things and cannot exist without them: materiality of properties directly derives from materiality of things.

According to Bunge (1977), people are unable to observe properties directly, and perceive them instead as *attributes*. Several attributes can potentially refer to the same property. The existence of an attribute does not imply that a particular property exists (e.g., the attribute *name* is an abstraction of an undifferentiated bundle of properties). While material things exist independent of an observer, individual observers may consider different attributes of things at different points in time. Indeed, attributes are basic abstractions of reality insofar as any attribute (e.g., color *red*, *roughness* of texture, *height* of a building) is a generalization formed by compressing diverse sensorimotor input (or memory) into a mentally stable coherent element²¹. Attributes are fundamental building blocks of representation to the extent that they can be used to identify instances and form higher-level abstractions (e.g., things with similar attributes can be grouped into *classes*). Properties can be *intrinsic* if they are inherent in things (e.g., *height* or *mass*) or *mutual* if they belong to more than one thing. The third principle states:

Principle 3: Attributes can be attached to an instance to describe its properties.

²¹ When considering visual modality, with every input interruption or environment change, such as movement of eyes (saccades) or of the object of interest, the focal object (stationary, or moving) is sensed differently by the retina, but operational constancy and equivalence of attributes, such as shape, color, length, texture, size are maintained (see, for example, Harnad 1990).

People use classes to group instances they deem equivalent in some way (see Fodor 1998; Murphy 2004; Smith and Medin 1981). According to Bunge, the equivalence is based on shared properties of things at a given moment in time. Classification allows humans to abstract from differences among instances, thereby gaining cognitive economy and ability to infer unobservable properties of things (Parsons and Wand 2008; Rosch 1978). For example, by stating something is a *bird* speakers can save the effort to communicate attributes they assume are true of birds (e.g., has heart, has feathers, probably can fly). Using classes improves the communicability and lessens the effort of having to provide an exhaustive list of attributes per instance. Classes are also intuitive when reasoning about instances. It is unnatural for users to refer to instance x in terms of its attributes alone. It is likely that users refer to x using some *class* (e.g., *dog*, *employee*, *bank*, *account*). Finally, knowing what classes users assign to instances reveals any biases in the kinds of attributes users attach to instances. The classes known to a person influence human perception, as illustrated by stereotype effects (Jussim et al. 1995) and categorical perception (Harnad 1990); knowing the classes users attach to instances, therefore, illuminates gaps and biases in the provided attributes. In summary, classes become a convenient and natural mechanism by which users can reason about instances and describe their properties of interest. They also help to understand the attributes provided. Finally, as demonstrated in Chapter 4, when given freedom to classify in an open-ended manner, non-expert users tend to provide classes (generally generic, "basic" classes) with high accuracy. Therefore classes are conceptualized as constructs that can be attached to instances.

Principle 4: A class can be attached to an instance to represent bundles of properties possessed by other instances described by the same class.

Despite the advantages, classes have a notable limitation. As discussed earlier, any two observers may fail to share the same class definition. In UGC settings, however, predicting how a particular user may understand a given class is challenging: “[c]lassifications that appear natural, eloquent, and homogeneous within a given human context appear forced and heterogeneous outside of that context” (Bowker and Star 2000, p. 131). For example, when two users “label” (the same) instance x as *employee*, it is unclear whether both users agree on attributes that define this class. For example, user 1 may consider employee to include part-timers and contractors, while user 2 may only consider full-time employees. In a UGC environment both perspectives may be valid, but it may be important to explicate each user's definition of the classes used.²² This leads to the formulation of the following principle.

Principle 5: Classes may be defined explicitly (e.g., in terms of attributes).

²² Whether it is necessary to make class definitions explicit may vary depending on context and classes used. For example, generally we may want to clarify the definition of a *bank account* or *planet* rather than more 'obvious' classes such as *human* or *rain*. This, however, depends on the target application: a paleontological or weather monitoring IS may be specifically interested in understanding how users define *human* or *rain*. Broadly, since all uses of data are infeasible to discover in advance, it is recommended to be explicit in class definitions.

The principles above can be summarized in a conceptual meta-model shown in Figure 2. It uses the proposed constructs of instances, attributes, and classes. As follows from Principle 2, instance is the main construct used to model UGC. An instance is manifested via one or more attributes. Since attributes cannot exist without instances, for an attribute to exist, it must be assigned to at least one instance. Attributes can also be used to form classes such that instances with shared attributes are considered members of the same class. A class, however, can also be attached to an instance directly, without having to specify the attributes – resulting in attributes being optional. Finally, as classes in UGC settings are attached to instances, no class can exist without an instance.

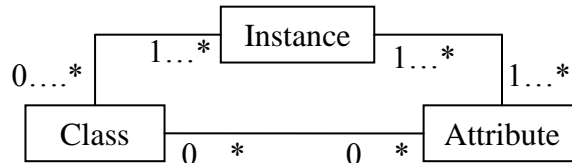


Figure 2. Instance-based meta-model

Following from the above principles, modeling UGC is based on representing particular *instances* via attributes, classes and interactions as perceived by particular users at certain moments in time. In contrast to representation by abstraction, the principles proposed above are founded on the assumption of *representational uniqueness* - each representation of the same instance may be different (i.e., expressed using different attributes and classes), including representations by the same user at different times. At the same time, representational uniqueness does not imply that every stored representation be unique, as two different users may independently provide the same set

of attributes and classes for the same instance; however in UGC environments all shared classes and attributes are difficult to determine in advance.

A consequence of representational uniqueness is the fact that capturing class-based abstractions *a priori* no longer becomes necessary. This deviates fundamentally from traditional conceptual modeling that guides analysis toward discovery and representation of domain specific class-based abstractions that capture commonalities among instances. This approach resolves the dilemma in modeling UGC uncovered in Chapter 4, whereby users were accurate when classifying at generic levels (which are not typically useful to the organizations), but using these levels engenders information (attribute) loss. In an instance-based representation, users can provide generic classes (e.g., bird) and then further describe the instance using any number of attributes (which, as Chapter 4 demonstrates, tend to be low-level, more specific, ones).

Representational uniqueness leads to IS development without relying on abstraction-driven grammars.²³ Under this approach, development proceeds by selecting a "flexible data model" and a "flexible user interface" (discussed in detail below). Users are then able to provide information according to their own conceptualization of reality

²³ This does not suggest that modeling is completely absent from IS development - it merely emphasizes the absence of a traditional specification of the classes of information that an IS is designed to manage. This thesis recognizes, however, that any development inherently involves some degree of modeling, a point considered in Chapter 6.

without having to conform to a particular structure. Such information can be stored in a flexible data model such as instance-based (Parsons and Wand 2000), graph (Angles and Gutierrez 2008), or semi-structured (Abiteboul 1997) data models. Several other promising schema-less databases have been proposed (Cattell 2011; Pokorny 2013).

For example, using the instance-based data model, information can be collected without having to classify relevant instances; information about instances can be stored in terms of attributes (Parsons and Wand 2000). Different users can supply different attributes for the same instance. Failure to agree on classes, relationship types or attributes is no longer problematic as any attributes and classes can be seamlessly captured. The attributes can be then queried to select instances stored based on classes of interest or other criteria. Thus, classes and other abstract constructs are not necessary before implementing such a system and conceptual modeling may not be needed for the design phase (at least not for the purposes of generating a database schema and other design elements).²⁴

²⁴ This chapter focused on the advantages of the proposed modeling approach. The limitations of this approach are considered in the Section 8.2 in the context of future research.

5.4 Chapter Conclusion

This chapter proposed principles of modeling intended to support development in UGC settings. With the growing importance of UGC, as exemplified by the case of citizen science, a pressing question is how to carry out conceptual modeling in this environment. Predominantly grounded in the realities of corporate settings, traditional conceptual models struggle to handle the diversities and uncertainties of the new environment. One consequence of modeling domains using the traditional modeling paradigm is decreased quality of information stored in these systems (as empirically demonstrated in Chapter 4).

In this chapter, I argue that modeling UGC should be to the extent possible driven by representation of (unique) instances rather than domain-specific abstractions. As a consequence, traditional activities performed during systems analysis (as described in Chapter 2), including creation of a global unified schema, no longer apply. Under this approach, development proceeds by selecting a flexible data model and a flexible user interface. Users are able to provide information on the instances of interest to an organization. Hence online contributors become free to provide information according to their own conceptualization of reality without having to conform to a particular structure.

The principles of modeling proposed here can be converted into testable propositions. For example, research can measure the impact of these principles on dependent variables of interest (e.g., domain understanding, problem solving, or information quality) (Topi and Ramesh 2002). This can be done by deriving IS objects based on the proposed principles and comparing them with those based on traditional

conceptual modeling. The principles can be further used to design IS or its components in a real (i.e., action design research) (Sein et al. 2011) or laboratory settings. The principles can also be used to evaluate existing conceptual modeling grammars or even suggest ways to develop graphic notations that could support communication and interaction during UGC IS development.

The next chapter further demonstrates the usage of the proposed principles by describing the development of an information system artifact - a real system designed to capture user-generated content. Chapter 7 employs the proposed principles to evaluate the impact of conceptual modeling on dataset completeness in a real citizen science IS.

6 Demonstration of the Principles for Modeling UGC in a Real Citizen Science Information System

To provide a "proof by construction" (Hevner et al. 2004; Nunamaker et al. 1991) and demonstrate the application (Gregor and Jones 2007; March and Smith 1995) of the proposed principles of modeling UGC presented in Chapter 5, I implemented the principles by re-designing a real citizen science IS, NLNature (www.nlnature.com). Exposing abstract principles via instantiation follows a general recommendation in the design science literature (Gleasure et al. 2012; Gregor and Jones 2007).

6.1 NLNature Background

The NLNature project was launched in 2009 by Dr. Yolanda Wiersma, a biologist at Memorial University, Canada, as part of a larger Canada-wide initiative (the Participatory Geoweb for Engaging the Public on Global Environmental Change) to investigate how to engage the general public with issues of environmental change by means of interactive communication technologies (Parfitt 2013; Sieber 2012).²⁵ The project is a partnership among leading Canadian universities, including University of British Columbia, McGill, University of New Brunswick, University of Calgary, Ryerson

²⁵ <http://rose.geog.mcgill.ca/geoide/>

University and Memorial University. The specific scientific objective of NLNature is creating an online IS to map biodiversity of Newfoundland and Labrador (a territory of over 150,000 square kilometers) based on amateur sightings of nature (e.g., plants, animals).

Investigating NLNature reveals challenges of conceptual modeling in UGC environments. I have been engaged in NLNature from the beginning (2009): first as an IT consultant and later, as a co-investigator. Typical to other design science research, the academic involvement was triggered by a real-world problem (Hevner et al. 2004) of representing unpredictable user input from non-experts with high veracity.

The project proceeded through two phases: class-based (2009-May 2013) and instance-based (May 2013 - Present). In the first, as no principles of conceptual modeling for citizen science existed (Lukyanenko and Parsons 2012), the project was developed using a traditional class-based approach to conceptual modeling. An evaluation phase began as soon as the project was launched and revealed limitations and negative consequences of approaching citizen science with traditional modeling. I then re-designed the project in 2013 to implement the proposed modeling principles.

6.2 Phase 1 Design

In Phase 1, the design strategy to ensure information quality and participation was informed by prevailing practices in online citizen science.²⁶ Traditionally the first step in conceptual modeling is to identify a set of concepts (entity types, classes) that describe the domain (Parsons and Wand 1997). Consistent with similar projects (e.g., www.eBird.org, www.iSpot.org.uk), the objective of data collection was positive identification of species. Consistent with traditional conceptual modeling, therefore the observed instances would be primarily classified as species-level classes.

Focusing on species-level classes was driven by the information requirements of the scientists - the sponsors of the project. Species are widely-established units of monitoring, international protection and conservation (Mayden 2002). This level of classification has been focal in broader citizen science research and practice (Dickinson et al. 2010; Parfitt 2013; Wiersma 2010). Major citizen science projects (e.g., eBird.org, which collects millions of bird sightings monthly) implement prevailing modeling approaches (e.g., Entity-Relationship) and collect observations of instances as biological species (Parsons et al. 2011; Wiggins et al. 2013).

²⁶ As this thesis is concerned with the impact of conceptual modeling on IQ, it focuses on conceptual modeling phase of the project and considers other phases only when relevant.

The project sponsors suggested a mixed convention of biological nomenclature and general knowledge (“folksonomy”) to conceptually organize entities about which information was to be collected. In this approach, species-level classes became lower-level classes in a generalization-specialization hierarchy where higher-level classes were intuitive ones (see Figure 3). Hence, if a user was to select the top-level class first (e.g., "Sea Bird"), this could limit the species-level options (e.g., to only sea birds) helping the user to locate the intended one.

Conceptual modeling was performed using the popular UML grammar (Dobing and Parsons 2006; Evermann and Wand 2006; Grossman et al. 2005; Jacobson et al. 1999). A relational database was designed based on the conceptual model (Teorey et al. 1986); the same model informed menu items and the options in the data collection interface (see Figure 3). To improve information quality, users were allowed to collaborate and assist each other in identifying species in a social-networking style (e.g., post comments, exchange emails) - a practice recommended by researchers in citizen science (Silvertown 2010). Additionally, verification mechanisms (e.g., location analysis and expert verification) were implemented.

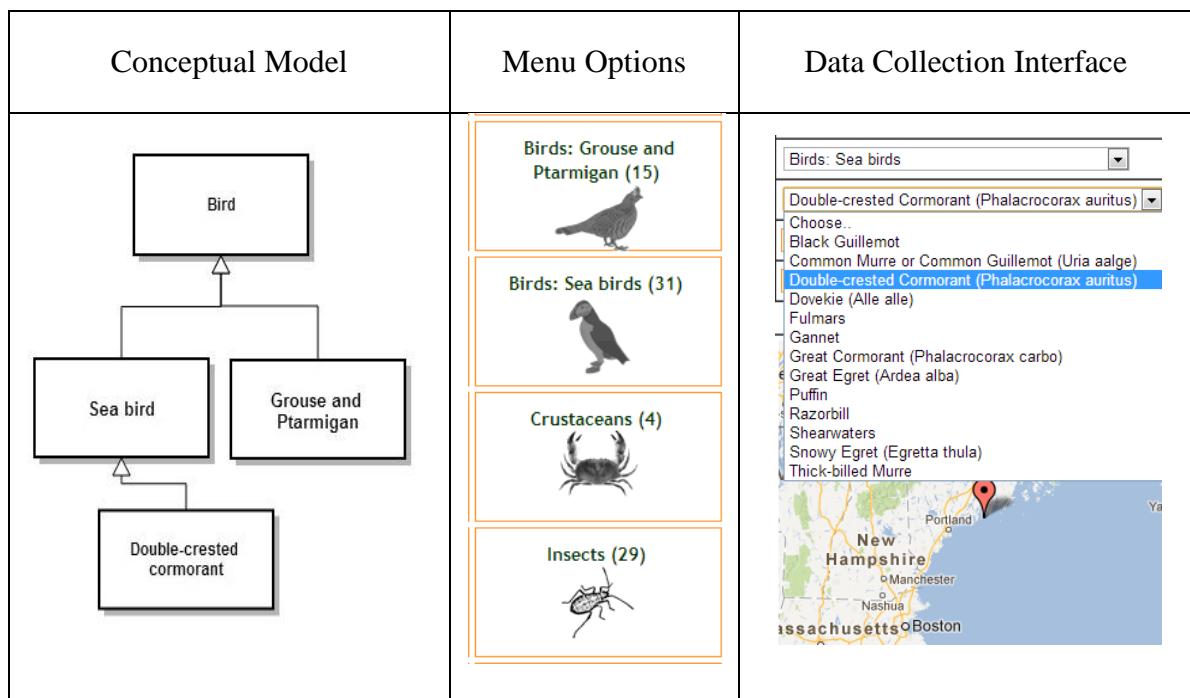


Figure 3. Conceptual model fragment and user interface elements based on the model in Phase 1 NLNature.

Once the project was launched, assessments of IQ were performed (including analysis of contributions, comments from users, and benchmark comparisons with parallel scientific sampling).²⁷ The project team (e.g., Kallio 2012) determined that the quality and level of participation were below expectations. Based on the arguments outlined in Chapter 3 of the thesis, I identified the class-based approach to conceptual modeling that supported the system as a detriment to both quality and participation. The

²⁷ A detailed discussion of the IQ issues on the Phase 1 NLNature is outside the scope of this thesis.

analysis of user comments suggested that some users, when unsure how to classify unfamiliar organisms, made guesses (to satisfy the requirement to classify organisms). A vignette with an observation classified as *Merlin* (*Falco columbarius*) where the observation creator admits to guessing is given in Figure 4. Notably, it took almost a year for another member to report an incorrect classification.

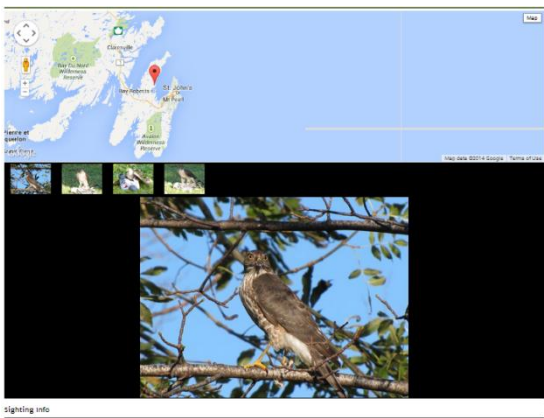
Screenshot of the observation	Public correspondence between the observation creator, Lynette, and another user, Timothy.	
	Lynette Nov. 17 2011	<i>I think this is a merlin... she (he?) killed a pigeon in my garden and ate breakfast right there, as the pigeon was too heavy to carry off...</i>
	Timothy July 28 2012	<i>Actually an accipiter. Sharpshinned hawk</i>
	Lynette July 28 2012	<i>Thank-you, Timothy! I'm an amateur, I Was guessing as to what it was!</i>

Figure 4. A vignette of an observation classified as Merlin (*Falco columbarius*) where the observation creator admits to guessing.

Additionally, in several cases, the organisms could not be fully described using attributes of the correctly chosen species-level class (e.g., morph foxes had additional attributes not deducible from the class *Red fox*). Finally, there was evidence that many observations were not reported because of the incongruence between the conceptual model and user views. For example, in contrast to biological nomenclature shown in

Figure 3, *Double-crested cormorants* may be considered by non-experts as *shore birds*, rather than *sea birds*, due to the strong association with shore areas; as a result a user may not be able to locate a *Double-crested cormorant* option under the *shore bird* level). The identified threats to information quality and user engagement motivated an effort to implement instance-based modeling on NLNature and, at the same time, provided an opportunity to offer an expository "proof of concept" of the proposed design principles in a real setting.

6.3 Phase 2 Design

IS development guided by instance-based modeling principles represents a fundamental shift from the traditional paradigm. Whereas traditional IS development begins with the elicitation and analysis of user requirements (Browne and Ramesh 2002; Jacobson et al. 1999), instance-based modeling suggests representation of individual (unique) instances. Consequently, although the project had access to a stable cohort of users - the scientists - I chose not to represent their views explicitly in a conceptual model (unlike in Phase 1). Instead, I elicited the intended **project objectives**, which included monitoring species distributions, informing conservation policy, protecting endangered species, and educating students and the general public. At the same time, the project was to be sensitive to the contributors' points of view and to the extent possible facilitate discoveries and unanticipated uses of data.

During the interviews with scientists, I identified the **domain of the project** to be all of natural history (i.e., plants, animals, and other taxa). Instance-based modeling according to the principles in Chapter 5 has no mechanisms to set domain boundaries - a

user may report an instance of a rock along with an instance of a bird. However, knowledge of the target domain can be leveraged in generating instructions to guide data collection to the potentially relevant (for the sponsoring organization) instances. Since NLNature's mandate was the provision of data to satisfy the sponsoring organizations' information needs, the IS design should remain sensitive to these views. However embedding these views in the *deep structure* of the IS (Wand and Weber 1990), such as the conceptual models and, consequently, database tables, would violate the representational uniqueness assumption. Consequently, I embedded organizational views in the surface structure (i.e., more mutable user interface elements) of NLNature. The organizational information requirements were reflected in the *data collection instructions* to accompany data collection fields and descriptions and explanations of the objectives and purposes of the project (e.g., see Figure 5).

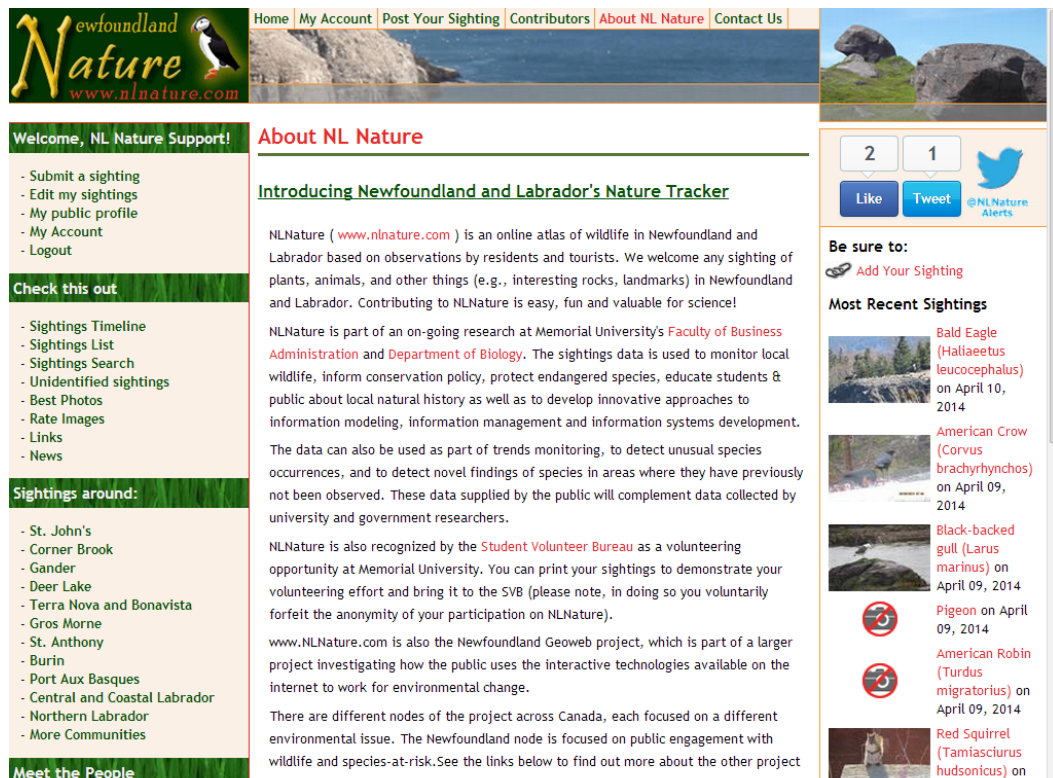


Figure 5. The "About Us" page on NLNature Phase 2 that describes project's focus.

Confining the organization's information requirements to surface elements constitutes a reasonable compromise between instance-based modeling and the pragmatics of projects driven by specific interests and agenda. As surface elements of IS, instructions and descriptions become mutable and can be refined without having to modify the deep structure. They also do not stand in the way of user expression (in contrast to traditional class-based structures when they are incongruent with data contributors' views), particularly if they make an explicit call for unanticipated kinds of instances.

Following the assumption of representational uniqueness, I did not engage in additional requirements elicitation to discover views of potential citizen scientists. Hence,

no consensus-building or view integration activities were conducted. A major part of IS development – the creation of a formal representation of knowledge in a domain - was a relatively minor phase - mostly aimed at understanding organizational needs to be reflected in surface elements of NLNature. Compared with Phase 1, following the instance-based principles significantly simplified systems analysis of citizen science and appeared to address the challenges of modeling UGC (discussed in Chapter 5).

Instance-based modeling advances the principle of representing instances and implies a schema-less database design. There has been increased interest in and development of flexible NoSQL databases providing several schema-less databases to store user input (Cattell 2011; Pokorny 2013). Potential candidate data models included key-value pair (DeCandia et al. 2007), document-focused (Chang et al. 2008) instance-based (Parsons and Wand 2000) and graph (Angles and Gutierrez 2008) data models. Of these, the closest model was instance-based (Parsons and Wand 2000) as it shares the ontological and cognitive foundations underlying this research and includes the relevant modeling constructs. Consequently, NLNature adopted the instance-based data model to store UGC.

The instance-based data model upholds the primacy of instances and assumes every instance may possess unique attributes (Parsons and Wand 2000). Classes are formed based on the principle that one can classify things based on a subset of their shared attributes. Since an instance can possess very many attributes, it can belong to a very large number of potential classes, depending on the context. Under the instance-based data model (Parsons and Wand 2000), users are not forced to classify instances

using predefined classes (such as biological species), which relaxes the constraint for non-experts to understand and conform to a chosen taxonomy. Using attributes makes it then possible to capture individual variations of organisms (addressing the issue of storing unique insights of contributors). The attributes can be queried *post hoc* to infer classes of interest (e.g., species).

An instance-based data architecture can be deployed on top of the popular and widely available relational database management software (Parsons and Wand 2013; Parsons and Wand 2000). To hold information about instances at a specific moment in time I created the "Observations" table (see Figure 6). The table contained date and time of the instance observation (guided by the assumption that instances are observed at some moment in time).²⁸ NLNature stored attributes and classes in a generic table "Concepts" that contained a unique identifier, a concept name, and a flag that distinguished classes from attributes. The "InstancesConceptsXref" held any attributes users provided for an instance containing concept identifier and instance identifier as foreign keys. The table "ConceptsXref" contained the primary key from the class or attribute and a primary key from a class or an attribute, thus making many-to-many relationships possible. For

²⁸ This thesis provides a simplified implementation. For example, in a real project like NLNature additional attributes may be included in each table, including a time stamp, system ID of the record creator, and any security, validation and monitoring keys. This information belongs to the *design rather than the application domain* and is outside the scope of this thesis.

example, *boreal felt lichen* could link to the following attributes: fuzzy white fringe around the edges, greyish-brown when dry, has red dots, leafy, and slate-blue when moist.

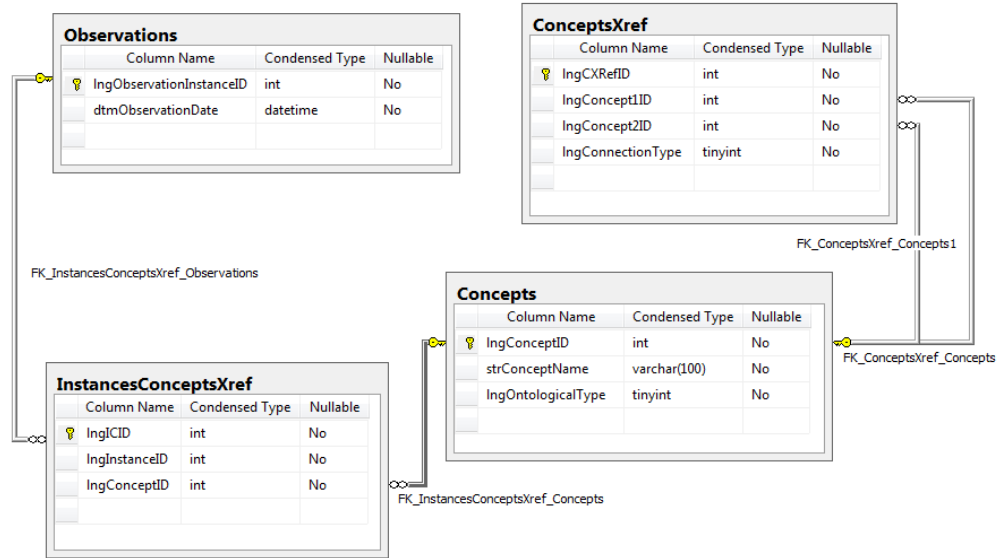


Figure 6. Logical view (table schema) of the NLNature's instance-based implementation

I then proceeded with the development of the user interface. As the proposed principles are mainly concerned with deep structure of IS, the database design was relatively unambiguously derived from the proposed modeling principles. In contrast there were challenges in developing a congruent user interface and other elements of the surface structure. Traditionally, surface elements of a system (such as a user interface, navigational structure, and menu choices) conform to structural assumptions at the deep (i.e., conceptual) level (Wand and Weber 2008). Since the proposed principles are founded on the assumption of representational uniqueness (discussed Chapter 5) by which different users may provide potentially unique attributes and classes, it followed that surface elements should support the variability of attributes and classes. At the same time,

no strategy for directly mapping the principles into surface-level design objects could be derived. As recommended in Newell and Card (1985), Arazy et al. (2010), and Kuechler and Vaishnavi (2012), I broadly surveyed relevant theories in psychology, human-computer interaction, software engineering and IS to seek additional, design-specific guidance.

As implied by Principle 2 in Chapter 5, the focus was on how to collect attributes and classes that describe instances. In traditional IS development, information collection is driven by the classification structure (and relevant constraints), in which case typical data entry may involve classifying the instance into one or more predefined classes (see, for example, the data collection interface from Phase 1 of the project in Figure 3). Modeling UGC involves managing information about instances in terms of potentially unique attributes and classes. Here, a practical question is how to choose interface elements compliant with the proposed principles. For example, a website could still present attributes and/or classes as a list, allowing users to select/check off applicable ones. One advantage of this approach is ease of interaction as users do not have to expend the effort in typing attributes and classes. This is consistent with the established menu-driven paradigm of user interface design (Newell and Card 1985).

At the same time, collecting instance information using menu-driven options appears incongruent with the proposed principles. In this implementation, all applicable classes and attributes would need to be modeled in advance - which violated the representational uniqueness assumption. Further, small screens on mobile devices make it difficult to present large amount of information (e.g., a long list of attributes) and could

impede user interaction (Ghose et al. 2012). There is also a concern regarding possible effects due to priming, ordering and cuing (Goldwater et al. 2011). For example, if the "correct class" is at the end of a very long list of classes, some users may fail to notice it and abandon data entry.

The representational uniqueness assumption suggests that data collection interfaces be, to the extent possible, open and flexible. Following popular practice on social media websites (e.g., Facebook, Twitter), search (e.g., Google) and citizen science projects (e.g., www.iSpot.org.uk), I decided to use a prompt-assisted ("autocomplete") text field. This allows a participant to begin typing a class or an attribute and a prompt dynamically shows recommendations based on the string being typed (see, e.g., Figure 7). This approach has advantages over a traditional constrained-choice mode (such as in Figure 3). As a text field is always initially empty, it mitigates any adverse ordering and priming effects. It also enables users to seamlessly enter new classes and attributes - without having to move elsewhere for this task.²⁹ Finally, as more people become engaged with social media, the dynamic text field is becoming a norm. Related to this,

²⁹ In developing NLNature, I was additionally interested in comparing the new version of NLNature with traditional one using field experimentation (Chapter 7). The dynamic text field was also compatible with traditional *input* HTML tag allowing for comparison between two systems.

Kluge et al. (2007) found higher user satisfaction with IS that implemented a dynamic text field experience.

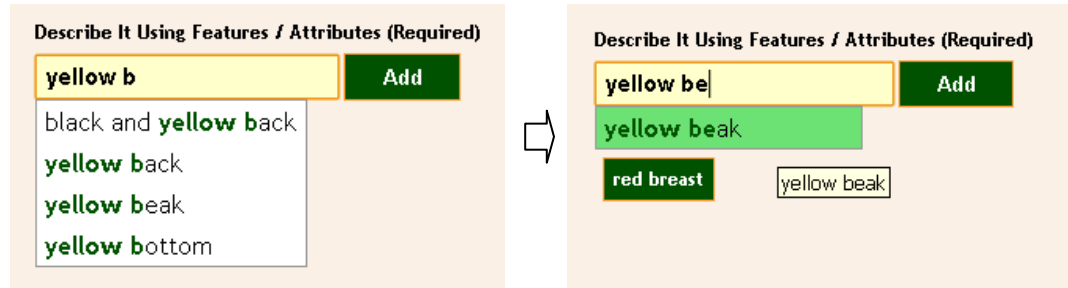


Figure 7. Example of data collection in Phase II

To guide participants to instances from the domain relevant to the sponsoring organization (as discussed earlier), a decision was made to instruct NLNature participants to provide attributes and, if possible, classes (see Figure 8). Since data collection based on instances was novel, detailed instructions for participants were provided on how to report observed instances. Specifically, immediately underneath the dynamic text field NLNature defined attributes:

Attributes (or features) are words that describe the organism you observed, including its properties, behavior and the environment.

The new interface also invites categories or classes if users are confident in classifying. When reporting attributes or classes, users were instructed to begin typing in the textbox and click "Add" or press "Enter" when finished. As soon as more than two characters are entered, a suggestions box would appear with the classes or attributes that contain the string entered. Users could select an item from the list or provide novel attributes and classes via direct entry. Once a user finishes providing attributes and

classes, the observation becomes public (optionally users may upload photographs). The website also contains a dynamic map on the front page of the project showing the most recent sightings (see Figure 9).

Type an Attribute or Category

Add >>

Instructions

Attributes (or features) are words that describe the organism you observed, including its properties, behavior and the environment.

- To add an attribute, type it in the box above and click "Add" or press "Enter" on your keyboard.
- You can add several different attributes, but please add one attribute at a time.
- You can add an attribute even if it is not found in the suggestions list. New attributes are welcomed!
- Please list as many attributes as you can.
- Once added, you can edit / delete an attribute by clicking on it.
- If you are confident about the identification, you can also add a category by typing it into the same box and clicking "Add".

Organism's Attributes and Categories

alone spotted white / gray dark eyes dark tip of

wings gray beak immature? in St. John's in the

ocean near shore near Signal Hill sporadic

movements swimming in water Gull Bird

Figure 8. NLNature Phase 2 data entry interface.

When using NLNature, users do not need to classify instances of interest (e.g., animals, lichens, geological forms) as would be required under traditional class-based designs. Instead, users provide attributes and classes of the observed instances. Different users can supply different attributes (or classes) for the same instance based on their knowledge. Failure to agree on classes or even attributes is no longer problematic as novel classes and attributes are accommodated.

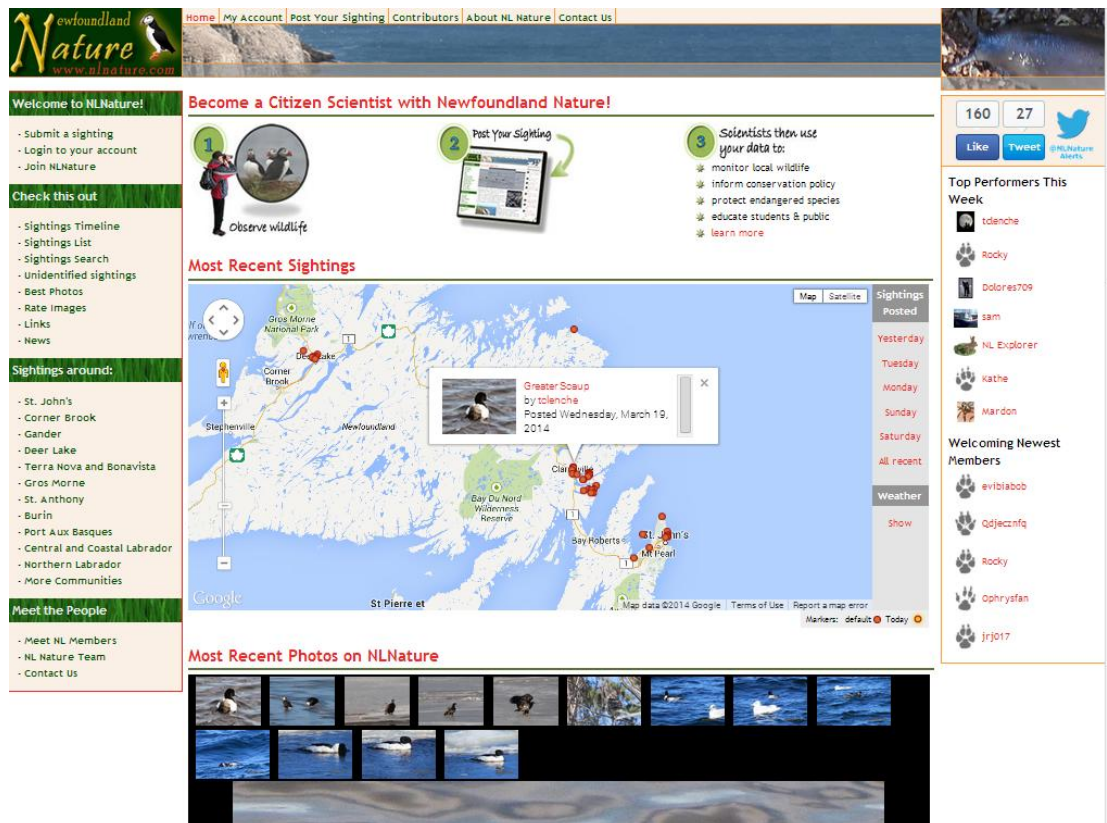


Figure 9. Redesigned front page of NLNature (public view)

By shifting the focus from a predefined classification to instances, modelers do not need to model a domain *a priori* in terms of the classes of interest. While NLNature based on the instance-based modeling principles may fail to deliver information in a predictable form to its sponsors, it opens novel opportunities for using this data in decision making. For example, scientists no longer need to create a complete specification of the kinds of instances assumed to exist in a domain. The openness of the IS itself should enable direct representation of novel classes - opening opportunities for discovery of new classes (Chapter 7 provides empirical evidence for this point).

The primary scientific object of an observation (the species observed) can be identified after the observation is recorded, provided the user reports enough attributes to produce a positive identification. When required, scientists can assemble a dynamic classification based on the collection of attributes that are of interest at a given moment. For example, if an attribute such as “behavior” is of interest, then at least two classes can be constructed based on values: nocturnal and diurnal animals. The same system can also use attributes that connect each species with a biological taxonomy to reproduce scientific biological classification. Thus, in principle, NLNature is capable of achieving the objectives of a traditional classification without the inherent limitations.

6.4 Discussion

The implementation of the proposed principles in NLNature has the potential to increase both the quality of citizen science data and participation rates. Unlike UGC projects that implement traditional approaches to modeling and assume a basic level of expertise from citizen scientists (e.g., eBird), NLNature allows for the full spectrum of contributors (Coleman et al. 2009) to participate. The value is that such data sets are generated by many “eyes on the ground;” thus, there is a higher likelihood of rare or unusual species being detected or for early detection of new trends. Hence, it is important to have a usable system that promotes a broad level of participation.

Instance-based NLNature represents a realistic compromise in citizen science. Non-experts do not always know the phenomenon that was observed. It is more realistic to expect a volunteer to remember some features of unknown species than to expect a precise classification and identification. Based on the premises of instance-based

modeling proposed in Chapter 5, this thesis hypothesizes that an IS developed by following these principles should result in high quality of UGC and greater user participation.³⁰

At the same time, it is important to note the implementation trajectory presented here is not the only possible one and other decision choices may be more fitting to the characteristics of the modeled domains. For example, this chapter does not specifically discuss a mechanism for tracing the identity of instances. In a vast space of natural history, it is difficult to identify identical individuals. The current implementation makes no explicit provisions for identifying two observed instances as the same. However, this can potentially be done indirectly, by computing similarity over the attribute space of the stored instances. For guidance, practitioners are advised to consult research on record deduplication in data quality (Batini et al. 2009; for review see Christen 2012; Madnick et al. 2009; Stoller 2009), data integration work in the context of schema matching (Batini et al. 1986; Doan and Halevy 2005; Evermann 2008; Heath and Bizer 2011; Lukyanenko and Evermann 2011; Sherman 2007; Spaccapietra and Parent 1994) as well as similarity

³⁰ To provide empirical evidence of the impact of modeling on dataset completeness I created an alternative version of the project (I decided to create a new version to ensure that any idiosyncratic features of the old version would not confound the results) following traditional class-based conceptual modeling. I then randomly assigned users to the "instance-based" and "traditional class-based" versions and analyze their performance in each condition. The results of this experiment are provided in Chapter 7.

theories in cognitive psychology (Gentner and Markman 1997; Goldstone and Medin 1994; Hahn et al. 2003; Holyoak and Koh 1987; Imai 1977; Mix 2008; Shepard 1962; Tversky 1977; Tversky and Gati 1982).

Another issue of interest is whether more advanced semantics should be captured in the NLNature database. As in the instance-based database can store any idiosyncratic attributes, the opportunity exists to increase both the number of provided attributes and their relevance to the sponsoring organization and data consumers by *guiding user input*. Such implementation can exploit semantic links between attributes. Psychology and ontology suggests that many attributes naturally correlate (e.g., *can fly* is highly correlated with *has feathers*) (Rosch 1978), form groups (i.e., those describing behaviour, appearance) leading to formation of sub-schemas (Murphy 2004) as well as precede other attributes (e.g., knowing that something *is blue* implies an attribute *has color*) (Bunge 1977; Parsons 2011). This information can be stored in a database, for example in the table "ConceptsXref" of NLNature, and be invoked for user input validation and guidance.

Links between attributes make it possible to support powerful inferences that can be leveraged in processing user input (e.g., by validating data entry or suggesting additional attributes to a user), and interpreting instance-based data. For example, if a user provides the attribute *has wings* then the system could probabilistically (e.g., based on prior observations and links in the "ConceptsXref" table of NLNature) infer it was a *bird*. It could also take advantage of property precedence (Bunge 1977; Parsons 2011) and ask for specific manifestation of *has wings*, such as *color of wings* or *size of wings*. Similarly,

once *has wings* is provided, NLNature can flag user input such as *lives in water* as inconsistent (and, if required, exclude it from scientific analysis). These design options while not immediately derivable from the principles proposed in Chapter 5, appear to be congruent with the principles and can be valuable extensions to NLNature.

6.5 Chapter Conclusion

NLNature provides an opportunity to demonstrate an implementation of the principles for modeling UGC proposed in Chapter 5. This not only provides a "proof by construction", but shows what aspects of IS development change by introducing the proposed principles in the development process. NLNature attests to the feasibility of the proposed principles and also provides a blueprint that practitioners can follow when developing UGC projects.

Modeling UGC following the instance-based principles promises to simultaneously leverage crowds to satisfy organizational information needs as well as harness creativity and unanticipated insights of the crowds. Indeed, both can be achieved as long as fundamental assumptions about information management change to better reflect the nature of UGC environments. By re-defining the fundamental unit to be instance (rather than class), crowd contributors with different levels of domain expertise and motivation can contribute relevant data.

NLNature's implementation of the proposed principles explicitly supports unanticipated uses when information about instances might be used for purposes not considered when a system was designed. For example, while scientists prefer *species* as a focal domain abstraction, instances of *oiled birds* may also become valuable (as they

might signal a potential environmental crisis), even if the precise identification at the species level is not provided. As argued in Chapter 3 and demonstrated in Chapter 4, information loss is inherent in class-based modeling. This implies that even correct species identification may not capture all attributes an observer may report. In contrast, the instance-based NLNature permits seamless capture of individual attributes (e.g., *appears sick, missing one antler*) generating information that would be challenging to capture using traditional modeling and enabling potentially more insightful analysis.

Using NLNature, this chapter provided an example of realizing the proposed principles in a real IS. In the case of NLNature, the analysis phase of IS development appears to be substantially reduced compared with traditional model-driven IS. Specifically, this chapter shows that it is possible to create an IS without a priori conceptual structures. There is considerable growth in the market of NoSQL databases leading to development of several popular commercial packages (Cattell 2011). Much of development in this area, however, has been driven by technical considerations, such as scalability, latency, and redundancy (Cattell 2011; Pokorny 2013). Considerably less attention has been dedicated to issues of conceptual modeling - a deficit that has been addressed in this chapter.

The next chapter evaluates the impact of conceptual modeling on dataset completeness by comparing the two versions of NLNature described in this chapter.

7 Impact of Conceptual Modeling on Dataset Completeness

Motivated by the findings from the three laboratory experiments in Chapter 4, Chapter 5 proposed a set of principles to model UGC. Chapter 6 demonstrated how a real IS can be developed that followed these principles. The implementation of the proposed principles in a real IS opens an opportunity to compare the impact on IQ of traditional, class-based modeling with the proposed instance-based modeling. This chapter investigates the effect of the two conceptual modeling approaches on *dataset completeness* using field experimentation.

7.1 Theoretical Predictions

Following the proposed definition of crowd IQ, data collection in UGC settings should be to the extent possible sensitive to the view of information contributors. Chapter 5 developed principles of modeling UGC that are congruent with the nature of UGC settings and abilities of information contributors. Traditional approaches requiring *a priori* classification (e.g., requiring users to select from a checklist of species) are usable only by more expert participants. As an alternative to class-based models, observations from citizen scientists can be collected and stored in terms of instances, their attributes and any classes that contributors deem relevant. This represents a more realistic approach to UGC. Non-experts online cannot always identify (or may not be willing to identify down to the level required by data consumers) the instance observed. It is more realistic to expect a volunteer to report some attributes of an instance than to expect a precise and accurate classification (which Chapter 4 showed to be highly unlikely).

The completeness of information stored in an instance-based IS can be compared with that stored in a traditional IS due to the fact that both systems represent *instances*. For example, records in traditional IS (e.g., such as those in popular citizen science projects), are about instances of interest reported in terms of the classes useful to project sponsors or data consumers (e.g., species). Instance-based IS can have records about the same instances, but their attributes and classes would naturally vary (reflecting different levels of domain expertise, motivation, and other contextual factors inherent in UGC settings). While in the latter case some classes and attributes relevant to data consumers maybe missing,³¹ information is still relevant insofar as it pertains to the instances of interest to data consumers, satisfying the definition of crowd IQ. Based on Proposition 3 (in Chapter 4)³², I hypothesize:

³¹ Collecting information without forcing it into a predefined class-based models poses a question of usefulness of the resulting instance and attribute data for the purposes that require species-level classification (which are common in biology). For example, can the attributes reported by non-experts be used by experts to reliably infer useful classes (e.g., species)? A positive answer would provide strong evidence of usefulness of data collected following the principles developed in this thesis. Investigating this question is beyond the scope of this thesis, but can be pursued in future research.

³² Proposition 3 (Dataset Completeness) states that class-based conceptual models undermine dataset completeness resulting in fewer instances stored when the classes defined in an information system do not match those familiar to the information contributor.

H-4.1 (Dataset Completeness). Contributors will report significantly more instances of biological organisms in the instance-based IS compared with an equivalent class-based IS.

Hypothesis 4.1 is primarily motivated by the contention that class-based modeling approaches may have inherent barriers to describing instances of interest. This undermines dataset completeness due to mismatches between the conceptualizations of online contributors and the class-based models embedded in the IS. Similarly, in rich and complex domains (e.g., science, healthcare, consumer markets), it may be difficult to determine in advance all relevant classes of things (regardless of whether or not participants are previously familiar with them). For example, projects may be local in scope (Sheppard et al. 2014) and concerned with monitoring and conservation in a small geographic area. Since distributions of plants and animals are not static, it may be difficult, if not impossible, to develop a comprehensive classification that can account for everything that may be observed in a given locality. Indeed, finding anomalies and outliers might be the *raison d'être* for some UGC projects. Even a single valid data point would spell success for a project like SETI@Home that leverages distributed crowd computing in search of extraterrestrial intelligence (Korpela 2012).

Organizations increasingly hope to harness UGC to learn something new about their target domains. One approach may be to encourage participants to contact the organizers when they encounter something unusual or not fitting into the predefined structure. Anecdotally, a discovery of an object previously unknown to astronomy, Hanny's Voorwerp, occurred when an online contributor, Hanny van Arkel discovered a

huge blob of green-glowing gas while performing a task of classifying galaxies in the project Galaxy Zoo (Lintott et al. 2009). The project schema could not accommodate this instance and van Arkel (sensibly) posted this information in a forum created to support the project. This post was eventually noticed by scientists. While online contributors may find workarounds to record information they believe is important, class-based IS lack inherent affordances to capture unanticipated attributes and classes. In contrast, instance-based information management is naturally suited for capturing any unanticipated phenomena.

This chapter compares the ability of two modeling approaches to capture unanticipated kinds of instances. To ensure equitable and conservative comparison, the focus is on *new* (i.e., previously absent from the project schema) *species-level classes* (as opposed to, for example, new attributes of instances). This comparison is conservative insofar as species-level identification is the explicit task in class-based IS and is arguably de-emphasized in the instance-based IS where the focus is on attributes and classes. Based on Proposition 3, I hypothesize:

H-4.2 (Dataset Completeness). Contributors will report significantly more instances of *new* (i.e., previously absent from the project schema) biological species in an instance-based IS than in a comparable class-based IS.

7.2 Method

To evaluate the proposed hypotheses, the experiment uses NL Nature (www.nlnature.com) - described in detail in Chapter 6. A field experiment offers several advantages. Using a real project allows tracking real user behavior (as opposed to

behavioral intentions). It also allows for real research participants rather than surrogates (e.g., students). Studying behavior directly is a growing trend in a variety of disciplines from economics to psychology, where scholars argue that deeper understanding of actual behavior and its circumstances affords unique insights about unobservable states of human mind (Bargh and Chartrand 1999). Conducting research in a real setting also increases external validity compared to similar studies conducted in laboratory environments.

Prior to the experiment, NLNature was in existence for four years. The project had 285 users who collectively contributed 788 observations - these sightings were made using a traditional (species-driven) user interface that was designed in accordance with prevailing approaches to citizen science (e.g., eBird.org). The low number of users and sightings were of concern and the project sponsor was looking to find ways to increase the number of observations reported.

The decision to conduct the experiment was made one year prior to its commencement in May 2013. The 12 months preceding the launch of the experiment were spent in planning and development. Importantly, all promotional activities were halted during this time to avoid attracting the attention of the public to the project and to ensure a fresh start for the new project. Preceding the launch of the redesigned NLNature, the activity on the website was low (see Figure 10). This allowed us to rebrand NLNature to the local community as a fresh start.



Figure 10. Traffic trend on NLNature before (prior to June 2013), during (June - December 2013) and after the experiment (December 2013 to March 2014).

During 2012 I substantially redesigned NLNature, changing its appearance and behavior (see the front page in Figure 11). The data collection interfaces were completely changed (see Figure 12). I also timed the launch of the experiment to coincide with the end of spring - the time when wildlife becomes accessible and people spend more time outdoors.

NLNature was promoted to the general public. I organized a series of community meetings in different parts of the province. In the week following the launch of the experiment I made a trip around the province covering over 3000 km, conducting 60 informal and 5 formal meetings. The website was also advertised/featured on local radio, television, newspapers, online (e.g., through Google AdSense and Search network, Facebook, Twitter, through website partnerships). All demographic groups were targeted in the promotional activities to ensure a sample of users representative of the members of the general public (rather than only keen naturalists). The project was coined as a local citizen science initiative in biology (with no details of the IS component of the project given). The call for participation invited anyone to participate, emphasizing that no expertise in biology was required. The promotional activities produced substantial traffic

in the project - peaking at 30,000 visitors per quarter at the height of the experiment in mid-summer of 2013 (see Figure 10).

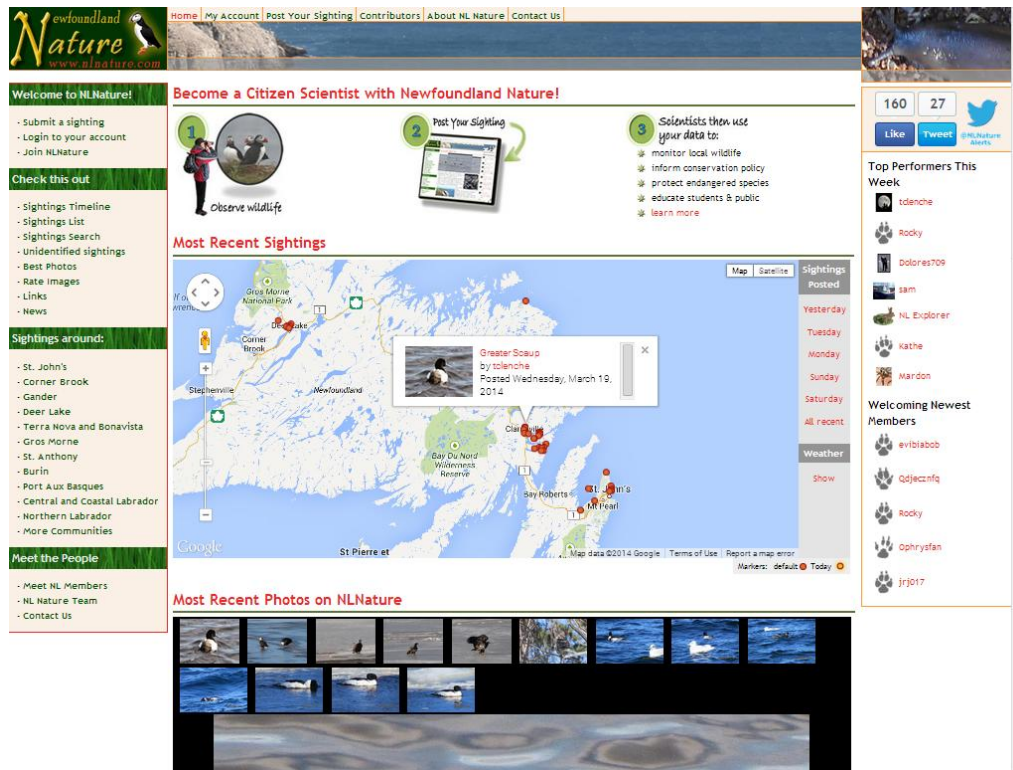


Figure 11. Redesigned front page of NLNature (public view)

To compare class-based and instance-based approaches to modeling, I used two different data collection interfaces, each corresponding to different conceptual modeling assumptions: class-based (species-level) interface and the instance-based interface (described in Chapter 6). The interfaces were designed to be visually similar and were dynamically generated from the same master template (differing only in the aspects relevant to the underlying conceptual modeling approaches).

Potential information contributors (citizen scientists) were randomly assigned to one of two data collection interfaces upon registration and remained in the originally

assigned conditions for the duration of the experiment. The data entry form required authentication to ensure that users were not exposed to different conditions. Regardless of the assigned condition, all users received equal access to other areas of the project (e.g., internal messaging system, forum) and equal support from the project sponsors. This ensured equivalent *facilitating conditions* (Venkatesh et al. 2003) across the three groups.

In the class-based condition, users were required to report sightings by selecting from a predefined list of species (see Figure 12). Since it is entirely possible that a contributor may not know or be confident in the species-level identification, the experiment provided an explicit option (with clear instructions) to bypass the species-level classification by clicking on the "Unknown or uncertain species" checkbox below the data entry field (see Figure 12). Following the principles for modeling UGC proposed in Chapter 5, in the instance-based condition NLNature instructed participants to provide attributes and, if possible, classes (see Figure 13). This allowed users to report sightings even if they could not determine a class for the instance observed.

Record your sighting: Step 2 of 3

<p>Type a Species (Latin or Common Name)</p> <input type="text"/> <input type="button" value="Add >>"/>	<p>Species Identification</p> <input type="text" value="Common Grackle (Quiscalus quiscula)"/>
<input type="checkbox"/> Unknown or uncertain species	
<p>Instructions</p> <p>Please identify the organism that you observed or select <i>Unknown or uncertain species</i>.</p> <ul style="list-style-type: none"> Type a Latin or Common name in the box above; we will then display a list of suggestions based on what you typed. Please select the appropriate option from the list. Then click "Add" or press "Enter" on your keyboard. Suggestions appear after at least 2 characters are typed. You may only add a species that appears on the suggestions list. If the species you observed is missing from our suggestions list - congratulations - you are the first one to post this species! Please select <i>Unknown or uncertain species</i> on this screen and write the species in the comments box (on the next screen). We will then add that species to the suggestions list and attach it to your sighting! 	

Figure 12. Class-based data entry interface

In both conditions, to see a list of options (classes or attributes) users were instructed to begin typing in the textbox and click "Add" or press "Enter" when finished. As soon as more than two characters are entered, a suggestions box appears with the classes or attributes that contain the string entered. In the class-based condition, participants were required to select an item from the list (or supply the new class in the comments, as per instructions). In the instance-based condition, participants could select an item from the list or provide novel attributes and classes via direct entry.

Type an Attribute or Category

Instructions

Attributes (or features) are words that describe the organism you observed, including its properties, behavior and the environment.

- To add an attribute, type it in the box above and click "Add" or press "Enter" on your keyboard.
- You can add several different attributes, but please add one attribute at a time.
- You can add an attribute even if it is not found in the suggestions list. New attributes are welcomed!
- Please list as many attributes as you can.
- Once added, you can edit / delete an attribute by clicking on it.
- If you are confident about the identification, you can also add a category by typing it into the same box and clicking "Add".

Organism's Attributes and Categories

alone

spotted

white / gray

dark eyes

dark tip of

wings

gray beak

immature?

in St. John's

in the

ocean

near shore

near Signal Hill

sporadic

movements

swimming in water

Gull

Bird

Figure 13. Instance-based data entry interface

As this chapter examines the context in which online contributors provide observations of natural history, the focus is on modeling phenomena in this domain. The conceptual model of the domain in the class-based condition, therefore, is a list of

species-level classes that reflects the intended uses of data by scientists (as discussed in Chapter 4). The choice of modeling only a single level in a classification hierarchy is driven by considerations of ecological validity as major projects (e.g., eBird.org) involve identification at a single, species-level.³³

The list of species was developed by an ecology professor - an expert in local natural history - when the project was first launched in 2009. It was deemed comprehensive as it represented most of the kinds of living things people are likely to encounter in the geographic area. NLNature became a live citizen science project in October 2009. During the four years preceding the current experiment, the list was updated periodically by the website members, who were encouraged to suggest new species (using the comments field available in the older version of NLNature). Biologists also reviewed the list periodically and updated it as needed. By the time the experiment began, the species list was stable with very infrequent updates and contained 343 species-level classes.³⁴

³³As this thesis is focused on the conceptual model of a domain, it ignores the database implementation details (i.e., how the conceptual model is translated into database tables). To ensure equivalence in query-write performance, entries in both conditions were written to the same database table "Concepts", as described in Chapter 6.

³⁴The list did not represent all species in the province, but deemed comprehensive for the kinds of things non-experts would be likely to experience (as decided by biologists).

As discussed earlier, the class-based version of NLNature implemented traditional approaches to conceptual modeling. When making specific design decisions (e.g., the design of data entry forms), it was important to have high ecological validity. Consistent with similar projects (e.g., www.eBird.org, www.iSpot.org.uk), NLNature instructed participants to provide a positive identification of species based on the predefined list. Following popular practice on social media websites (e.g., Facebook) and citizen science projects (e.g., www.iSpot.org.uk) I decided to provide options via a prompt-assisted text field. This allowed a participant to begin typing a class and a prompt would dynamically show recommendations based on the string being typed. As more people become engaged with social media, the dynamic text field is becoming a norm for data entry. It also appeared as a superior alternative to a dropdown list (such as in Figure 3; Chapter 6) as it mitigated potential adverse ordering and priming effects (see Chapter 6 for more discussion).

When designing the instance-based condition, much of the previous experience with class-based conceptual modeling, database normalization, and user interface design, could no longer be leveraged. Traditionally, surface elements of a system (such as a user interface, navigational structure, menu choices, code objects) conform to structural assumptions at the deep (i.e., conceptual) level (Wand and Weber 2008). Since the underlying conceptual and data model is instance-based, surface elements need to follow the instance-based principles. At the same time, no strategy for mapping the instance-based modeling into specific design objects is evident in the underlying theory (see Chapter 6 for more discussion). I used the implemented traditional condition as a template

for the instance-based one, as it was more important to ensure equivalence across conditions than produce the most effective implementation of the instance-based IS.³⁵ Specifically, I reused every design element included in the traditional condition that was not pertinent to the principles of the instance-based modeling. As with the class-based data entry, the interface began with the instructions but asked users to describe instances and attributes rather than classes. The same dynamic text box was used for data entry. To ensure equivalence across conditions, NLNature also provided users in the instance-based condition with options to choose from once they began typing. The options were based on a list of common natural history attributes (e.g., can fly, yellow beak) compiled before the start of the study. Unlike the class-based condition these options served as a guide and an example: users in the instance-based condition were not constrained to the predefined choices and were free to provide their own attributes and classes.

7.3 Results

Hypothesis 4.1 predicts that users in the instance-based condition will report more instances than users in the class-based (species-only) condition. Hypothesis 4.2 predicts that a greater number of instances of new species will be reported in the instance-based

³⁵ As the instance-based IS was expected to outperform the class-based IS, it was not necessary to design the most effective instance-based IS (principles of design were discussed in Chapter 5). Instead, the goal was to make sure that a valid comparison could be made.

condition than in the class-based condition (due to the inherent challenge of predicting all relevant classes in a citizen science environment).

The results are based on a six month period of usage, from June to December 2013. This period spanned low and high tourism seasons in Newfoundland and Labrador (peaking in late summer). It also allowed participants to observe major changes in ecology due to seasonal changes. The period corresponded to late spring, summer, fall and early winter in Newfoundland and Labrador.

In designing the project I followed established practices of engaging citizen science participants in scientific research including voluntary and anonymous participation (Robson et al. 2013; Snäll et al. 2011). In order to use NLNature, participants were required to accept a consent form that outlined the nature of their interaction with the website. Failure to accept the consent disallowed people from using any data-collecting features of the website. No incentives for participation were provided. Participation was voluntary. There was no requirement to stay on NLNature for any particular length of time or to submit sightings. Participants could provide as many sightings as they wanted and could quit using the website at any time without having to give the reasons. The instructions stressed that there was no requirement to provide some “minimal” amount of information even if the consent was accepted. Participation was anonymous and as a result no personally-identifying information was collected on NL Nature. The nature and presence of the manipulation was not disclosed to the participants.

The results of the study are based on the information provided by the website members who accepted the consent form after June 1st 2013 when the two manipulations

became live. Since June 2013 158 members accepted the consent form and were assigned to the two manipulations. Upon accepting the consent form, each user was randomly assigned to one of the two the study conditions. Since users could not be uniquely identified, their identification was based on the IP addresses. To prevent people potentially living or working in the same place from appearing in different conditions (with could contaminate the sample by making some people aware of different manipulations), users that shared the same IP address were always placed in the same condition.

In total, 79 participants were randomly assigned to the class-based condition and 79 were assigned to the instance-based condition. Some participants registered, but never landed on the observation collection page and hence were not actually exposed to manipulation (this was determined by analyzing server logs). The final number of participants who at least once visited the observation collection interface was 42 in the species-only condition and 39 in the instance-based condition. The remainder of the analysis is based on the information that was provided by these users.

While NLNature did not require users to provide demographic data, some volunteered this information by filling in an optional form. Fifteen participants indicated their age (50.9 avg., 15.54 st. dev). Seventeen participants indicated how many years they lived in Newfoundland and Labrador (18.9 avg., 17.30 st. dev). Fourteen participants provided number of hours per week they spend outdoors (19.1 avg., 15.54 st. dev.). Of the 27 people who provided information about their sex, 13 were female. While the majority

of participants abstained from contributing demographic information, those who provided information appeared mature with considerable local experience.

7.3.1 Hypothesis 4.1: Number of instances stored

To evaluate H-4.1, I analyzed observations provided by 81 participants exposed to manipulation in the two conditions. Since in the class-based condition a contributor might not know or be confident in the species-level identification, the interface provided an explicit option (with clear instructions) to bypass the species-level classification by clicking on "Unknown or uncertain species" checkbox below the data entry field (see Figure 12). The class-based interface further instructed participants to indicate in the comments box any class to which they believed the instance belonged. Since in this case a user could provide classes at levels other than species, such non-species observations were removed from the count for users in the class-based condition. Finally, since this thesis defined crowd IQ (Chapter 2) as the extent to which stored information represents the phenomena of *potential interest* to data consumers, I counted an observation as valid if it described *an instance* in the domain of biology (i.e., a living thing).³⁶

³⁶ This led to the removal of one observation of an island (although in-line with the use-agnostic IQ, even this observation may be useful at some point in time).

Table 10 reports the number of contributions in each condition, consisting of sightings made in the instance-based condition and species-level classifications in the class-based condition.

Table 10. Number of observations by condition

Experimental Condition	No of users in condition	Observations				
		Total	Mean	St. dev.	Skewness	Kurtosis
Class-based	42	87	2.07	2.56	2.08	4.23
Instance-based	39	390	10.00	37.83	5.47	29.66

Before proceeding with hypothesis testing, the assumption of normality in the data was tested using Shapiro-Wilks test. In each condition, the distribution of observations by user significantly deviates from normal (with $W=0.690$ and $p\text{-value}<0.000$ for the class-based and $W= 0.244$ and $p<0.000$ for the instance-based condition), due largely to the presence of outliers in each condition.³⁷ As seen from Table 10, in both cases the distributions are skewed and leptokurtic. This was confirmed using Kolmogorov-Smirnov

³⁷ By convention data points are deemed outliers if they are 1.5*interquartile range above the third quartile or below the first quartile (Martinez et al. 2004). The following frequencies of observations per user are outliers in the instance-based condition: 236, 39, 21 and 19 and 12 9, 7, 7, 6, 6, 5 and 4 in the species-only condition. I also verified that the most extreme value is a significant outlier using Grubbs' test, which confirmed that in each condition the extreme value (236 and 12) is a significant outlier (at 0.01 level).

and Anderson-Darling goodness-of-fit statistics where best fitting distributions were power-law, lognormal and exponential. Commonly, these are referred to as "long tail" distributions. Indeed, the top 4 contributors in the instance-based condition (or 10% of the user sample) produced 80.8% of the observations in that condition (in contrast, the top 4 contributors in the class-based condition produced 37.9% of the observations in that condition). These results are not surprising: long-tail distributions have been observed consistently in other user-generated datasets, including citizen science projects (Lukyanenko and Parsons 2013). The instance-based condition has greater mean, variance, skewness and kurtosis than the class-based condition (see Table 10). Figure 14 further illustrates this by showing that users in the instance-based condition tend to contribute a higher number of observations and few users in this condition contributed one or zero observations.

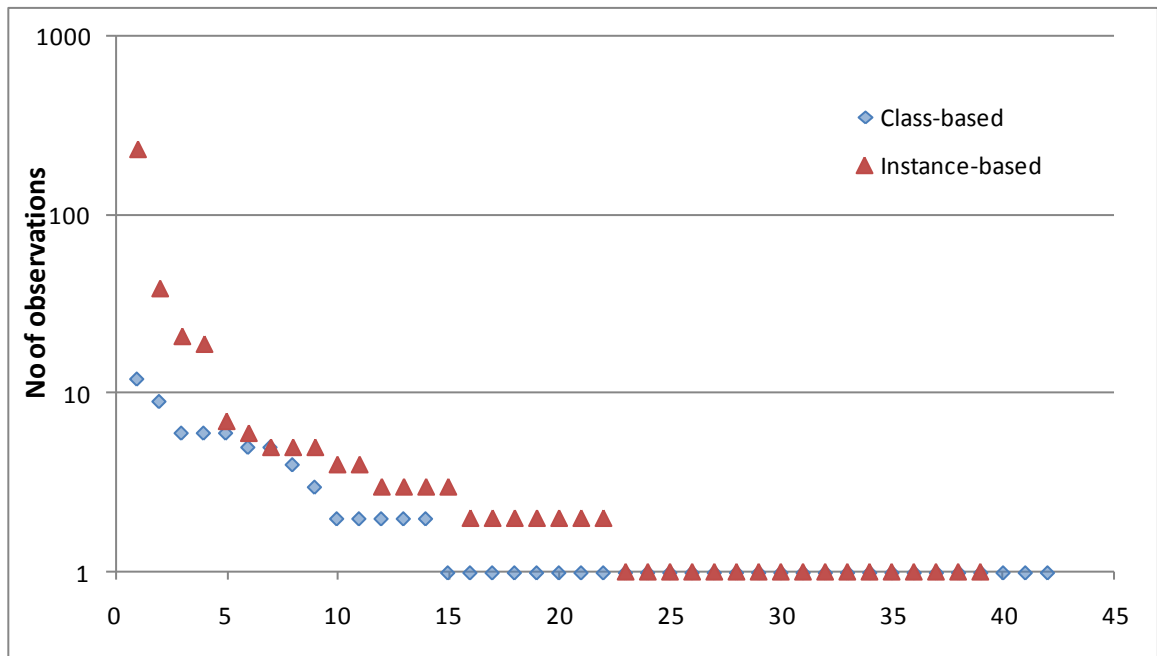


Figure 14. Number of observations per user in the two conditions

To determine if the difference in the number of observations per user is significantly different across the conditions, an exact permutation test was performed (Gibbons and Chakraborti 1992; Good 2001; Hayes 1996). The test samples from all possible outcomes without replacement to determine the exact probability of obtaining the observed difference. The permutation test can be performed if values in the two samples can be exchanged - meaning that users in both conditions could theoretically provide the same number of observations (i.e., the samples are comparable in principle, which is fundamental to testing differences in samples). Unlike other methods (e.g., parametric statistics, bootstrapping), assumptions about data distribution or population parameters are significantly relaxed, making the permutation test very general (Gibbons and Chakraborti 1992; Good 2001; Hayes 1996). The exact permutation test is suitable when data is not normally distributed, sample sizes are low and medium, outliers and ties (i.e., same values in two samples, as in Figure 14) are present. This test is preferred over approximations, such as bootstrapping that relies on permutation with replacement (Good 2001).³⁸

Based on the exact permutation test of observations per user between the two conditions, the p-value is *0.033*, indicating that users in the instance-based condition

³⁸ The permutation test is becoming popular and is being increasingly recommended with the availability of the requisite computational power (Hayes 1996).

provided significantly more observations than those in the species-only condition. This supports Hypothesis 4.1 and accords with the contention that different conceptual modeling approaches may result in significantly different numbers of instances of interest captured in IS.

To gain a deeper insight into the impact of modeling on information completeness, I further analyzed the categories and attributes provided to identify specific causes of lower performance by the users in the class-based group. Specifically, three (observable) behavioral patterns of users in the class-based condition led to lower information completeness. Below I elaborate on each pattern.

This thesis argued that since the class-based models constrain user input to predefined classes and attributes, users may not be able to record instances unless they provide classes that are congruent with the predefined structure in an IS. Evidence for this comes from the analysis of classes users entered in the dynamic textbox. The use of a dynamic textbox for data entry allows comparing words and phrases users attempted to submit against the classes defined in the IS. Whereas in the instance-based condition entering directly new attributes and classes was allowed, in the class-based condition the entries were vetted against the active species list and only matching ones were allowed (unless a user explicitly bypassed this step to report an unknown or new species).

Table 11. Examples of user input in the class-based condition that did not fit the species level of classification

Original user input	Reason for exclusion
Harvestman	Non-species
Slug	Non-species
Harbour Grace Island	Not on list; not animate
Otter	Non-species
Spider	Non-species
Hawk	Non-species
Black bear scat	Non-species
Toad	Non-species
Earwig	Non-species
Dolphin	Non-species
Caterpillar	Non-species
Soapberry	Non-species

The analysis of user input reveals instances of mismatch between the intended classification and the active class base (see Table 11). While NLNature specifically instructed users to provide species-level responses and identification at that level, as the prevailing practice in natural history citizen science, users still attempted to provide classes at other levels. These were generally at higher levels in the classification hierarchy (e.g., dolphin, toad, slug) potentially reflecting classification uncertainty (e.g., due to conditions of observation), and/or lower levels of domain expertise (non-experts are generally more comfortable with more general taxonomic levels).

Each case where the class provided in the comments box did not match the target (species) level was not included in the analysis above, contributing to the lower number of observations in the class-based condition. The existence of cases where users attempted to enter data at levels above the species-level provides evidence for the mismatch between the model of the contributor and the data consumer-oriented view embedded in the IS. This accords with the empirical findings in Chapter 4.

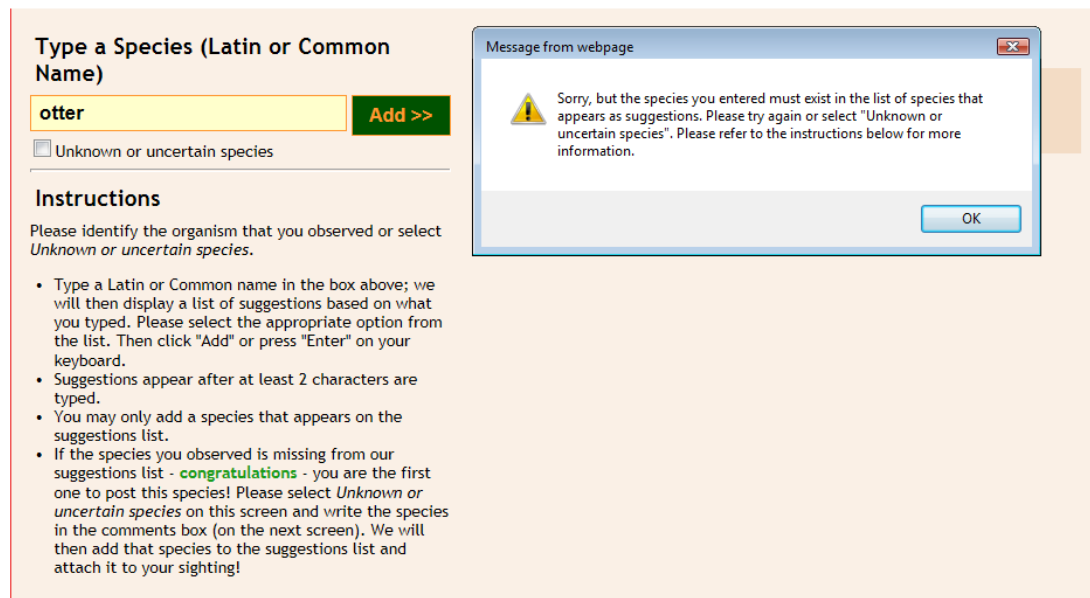


Figure 15. Feedback users received in the class-based condition when the word entered was incongruent with the classes defined in the model (notably, the message suggested bypassing classification as an option).

The second pattern observed showed that, when facing a structure incongruent with their own, some users changed the original submission. In several cases this resulted in loss of instances. For example, in one case a user began with typing "otter" (non-species level) - the entry was rejected by the system (listed in Table 11; see Figure 15 for a screenshot of the message the user received in this situation). The user then proceeded to record "Little Brown Bat (*Myotis lucifugus*)" instead (see Figure 16). In another case a user typed "grackle" (non-species level) 5 times before finally selecting "Common Grackle (*Quiscalus quiscula*)". A similar sequence occurred when a user first entered "toad" and then selected "American Toad (*Bufo americanus*)", "moose" and then "Moose (*Alces alces*)", "Canada loon" and then "Common Loon (*Gavia immer*)". Another user began with "black bear scat", and after two attempts to record it, typed "Black Bear

(*Ursus americanus*)". In all examples above the original input had to be changed by users to comply with the model. In the case of "otter" the instance of it was not stored.

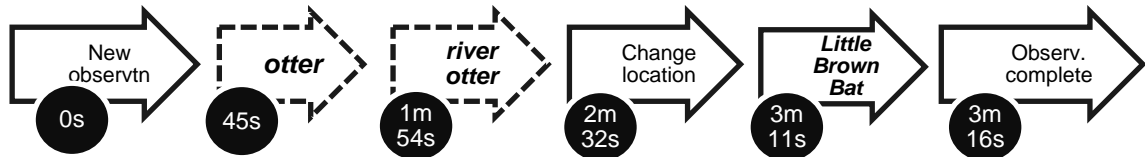


Figure 16. A timeline of the observation showing the loss of an otter instance ("otter" and "river otter" classes were rejected - shown in dashed lines - leading the user to modify the location of the sighting and report "Little Brown Bat" instead).

This chapter predicted that, when faced with unfamiliar classification structures, users may devise a workaround to record information. As the opportunity for direct entry is not provided, loss of instances may result. The data offer some evidence for this. In 12 cases, users in the class-based condition selected to by-pass species identification, but then failed to provide any species-level labels. These cases were also excluded from the final count of observations in the class-based condition.

Table 12. Examples of the basic-level categories provided in the instance-based condition.

Basic-level category	Reported frequency
Fly	29
Spider	18
Mushroom	17
Mosquito	8
Butterfly	7
Beetle	6
Bird	6

Another source of difference between the conditions is the prevalence of non-species-level classification in the instance-based sample. Many classes provided in the instance-based condition were at levels higher-than the species. Of 390 observations in the instance-based condition, 179 (45.9%) were not classified at the species level. For these observations, participants provided 583 classes and 69 attributes (222 distinct classes and 43 unique attributes). Among the classes provided, 110 were basic-level categories (see Table 12). As discussed in Chapter 4, basic-level categories are widely accepted in cognitive psychology as the generally preferred classification level for non-experts (Corter and Gluck 1992; Eimas and Quinn 1994; Markman and Wisniewski 1997; Rosch et al. 1976; Tanaka and Taylor 1991). The results from Chapter 4 further suggest basic level as a marker of low domain expertise. The reporting of basic-level categories can stem from at least three (possibly overlapping) sources:

- (a) low level of domain expertise of some users, as argued in psychology literature and as demonstrated in Chapter 4;
- (b) conditions of an observation (e.g., too dark, fleeting, at a distance) when positive identification at more specific levels could not be made;
- (c) attempts to provide additional evidence in cases when confidence in species identification is low.

The obtained results demonstrate that the mismatch between the conceptualization by users (situational or expertise-related) and those embedded in the IS contributed to the lower number of observations in the class-based condition. This supports Hypotheses 1.

7.3.2 Hypothesis 4.2: Number of novel species reported.

Hypothesis 4.2 posits that a greater number of new species would be reported in the instance-based condition than in the class-based condition. Users in both conditions provided 997 attributes and classes including 87 in the class-based and 910 in the instance-based condition (see Table 13). Of these 701 attributes and classes were new - they did not exist in the system prior to the experiment and were suggested by users as additions. This was done directly by users in the instance-based condition and indirectly (via comments to an observation) by users in the class-based condition.

Table 13. Number of observations and categories and attributes by condition

Experimental Condition	No of users in condition	Classes and attributes				
		Total	Mean	St. dev.	Skewness	Kurtosis
Class-based	42	87	2.49	2.62	1.97	3.49
Instance-based	39	910	26.00	117.14	5.36	19.8

During the experiment, 126 new species-level classes were suggested by the participants - 119 in the instance-based and 7 in the class-based condition (see Table 14). In each condition, the distribution of new species by user significantly deviates from normal ($W = 0.430$ and $p\text{-value} < 0.000$ for the class-based and $W = 0.232$ and $p < 0.000$ for the instance-based condition). The distribution is long-tailed in the instance-based condition (fitted using Kolmogorov-Smirnov and Anderson-Darling goodness-of-fit) and uniform ($\text{Chi-squared} = 47$, Monte Carlo $p = 0.424$) in the class-based condition. Based on the exact permutation test, the number of new species is significantly greater in the

instance-based condition ($p=0.007$), providing support for Hypothesis 4.2. This suggests that instance-based approach to modeling may be more effective for capturing data about unanticipated phenomena of interest.

Table 14. Number of new species reported by condition (repeated sightings excluded)

Experimental Condition	No of users in condition	New Species				
		Total	Mean	St. dev.	Skewness	Kurtosis
Class-based	42	7	0.17	0.44	2.53	5.96
Instance-based	39	119	3.05	13.17	5.35	28.51

Users also provided interesting attributes for some sightings. As implied in Proposition 2 (Chapter 3), these attributes offered *additional* information not inferable from the classification labels attached to instance:

- attributes describing situational behavior of the instances observed (e.g., *mating, hopping, fluttering together*);
- attributes describing something unusual about an instance (e.g., *tagged, only has one antler*);
- attributes describing the environment / location of the instance (e.g., *near highway, 10 feet away from highway, near bike trail*).

As these attributes cannot be predicted from simply knowing the species (e.g., while *moose* are known to *appear near highways*, one cannot conclude that the observed moose was near a highway if this information is not explicitly provided), they constitute information beyond what would be normally collected in a traditional class-based model.

Thus, unless appropriate designs are provided to seamlessly capture these attributes, they can be potentially lost. The field evidence of potential information loss provides additional support for the findings obtained in the laboratory setting (reported in Chapter 4).

Interestingly, several sightings of biological significance were reported during the experiment. These included unanticipated distribution of species (e.g., vagrant birds, fish and insects), a mosquito alien to the geographic area of the study³⁹, and a discovery of a possibly new species of wasp (presently pending scientific verification). All these occurred in the instance-based condition. It is also notable that some of the new organisms suggested by the instance-based users belonged to classes that were poorly represented in the project schema of the original class-based condition, including microorganisms and insects (e.g., 29 sightings of flies, 10 sightings of moth, 8 sightings of mosquitoes).

³⁹ *MUN Science News [Oct 4th, 2013]*: Citizen scientist detects sighting of mosquito thought to be carrier of West Nile <http://www.mun.ca/science/news.php?id=2579>

7.4 Discussion

The results of the field experiment demonstrate that modeling approaches affect dataset completeness and add to the evidence of the impact of conceptual modeling on IQ provided in Chapter 4.

Using a real IS project - an online natural history citizen science website, www.nlnature.com - this field experiment found that participants provided on average more observations when assigned to the version that implements instance-based, rather than the traditional, class-based modeling. Similarly, participants in the instance-based condition provided a greater number of novel classes of organisms. The results indicate that traditional modeling presents a barrier to providing information that appears to be mitigated by the instance-based modeling.

It is also notable that of the top 5 contributors, 4 belonged to the instance-based condition - collectively producing 315 sightings - 80.8% of the observations in the instance-based condition and 66.0% of all the observations collected during the study period. In contrast, the top 4 contributors in the class-based condition created 33 observations - or 37.9% of the observations in their condition and 6.9% of all observations. Although too small for statistical significance testing, this suggests that instance-based modeling might encourage the rise of "superstars" - people who contribute a disproportionately large share of the projects' content. Given that the typical distribution of user activity in UGC projects is long-tailed, superstars constitute a stable core of the project - a group of regular and potentially most loyal users. Nurturing the growth of superstar users may be central to a project's success, as they play a key role in content

production, dissemination of ideas and influencing other people (Chau and Xu 2012; Zhang et al. 2013).

Instance-based conceptual modeling appears to be more effective at capturing unanticipated phenomena. Users in the instance-based condition reported 17 times more observations of *new species* than in the class-based condition. One concern about the definition of information completeness from the perspective of data creators is that this may result in information that is irrelevant and of no value to the sponsoring organization. The findings appear to point to the contrary. Instance-based users outperformed the users in the class-based condition in the task with the predefined focus on species. A potential explanation for this paradox has to do with the increased flexibility and freedom afforded by the instance-based model. While the class-based users were given mechanisms to report new species, it was not direct and seamless. In several instances, users in this condition appeared on the path to provide new classes (by clicking on the bypass identification button), but contrary to the instructions, provided no valid descriptions in the comments. Another reason for the lower number of new species in the class-based condition might be related to the fact that users in this condition were directly exposed to the schema of the project - and thus could have formed a preconceived notion of the kinds of things that were of interest to the project sponsors. Indeed, users in that condition were *required* to select from predefined options. In contrast, users in the instance-based condition were *not required* to comply with any predefined options. Notably, some of the new instances logged by the instance-based users belonged to groups that were originally poorly represented in the project schema, including spiders, flies, and mosquitoes. These

organisms are readily observable by all users, but were nevertheless reported extremely rarely in the four years preceding the experiment even though the project explicitly embraced "all natural history". A widely-held assumption in citizen science holds that non-experts mostly report "charismatic" organisms, fueling concerns that citizen science produces a distorted view of biodiversity (Boakes et al. 2010; Galloway et al. 2006). The results of this study indicate that the imposition of a schema may bias participants toward predefined options and the bias may be mitigated using instance-based modeling.

Despite finding significant differences between the two conditions, it is notable that, among the information provided by participants in the instance-based condition, many classes were at the species-level of granularity. Such level of granularity is natural for domain experts, whereas novices are more comfortable with the more generic classes (as demonstrated in Chapter 4). This indicates that, despite efforts to attract members of the general public, many participants on NLNature might have had higher-than average levels of domain expertise. This may be explained by the fact that, being unaware of the novel experimental condition, prospective novice participants might have assumed that getting engaged in the project required some level of domain expertise. This could have dissuaded non-expert participants from joining and discovering the instance-based condition.

While participants in the instance-based condition provided more observations than participants in the class-based condition, a natural question arises as to the extent to which the instances in the instance-based condition belonged to classes (species) provided on NLNature before the start of the experiment. This question is important as these

classes typically support intended uses of citizen science information by the scientists. In the instance-based condition, participants provided 51 of the 343 (14.9%) species that were in the schema of the NLNature before the start of the experiment. By comparison, in the class-based condition participants provided only 36 (10.5%) of the original species. While this may be in part due to the overall larger number of observations in the instance-based condition (there were 390 observations in that condition and only 87 observations in the class-based condition), it illustrates that the use-agnostic instance-based approach does not necessarily result in failure to capture information known to be of immediate relevance and usefulness to data consumers.

There are several limitations of the presented field experiment. One general concern relates to the nature of empirical evidence obtained as a result of field experimentation. While using field experimentation offers advantages (discussed earlier) the results should be interpreted with caution. Working in a field setting raises common concerns about experimental control. One issue is ensuring that users in one condition were not experiencing treatments in different conditions. I tried to address this by using password authentication before any manipulation could be experienced. Having to enter (and remember) user name and password, however, potentially deterred some (e.g., less determined) users from engaging.

Another issue is whether the users of NLNature were representative of the broader population. In conducting the experiment, considerable effort was made to reach as many different segments of population as possible (as expounded above). At the same time, the analysis of observations revealed an unexpectedly large proportion of species-level

identifications - indicative of domain experts (Tanaka and Taylor 1991). This can be potentially explained by the volitional nature of the project where users with domain knowledge or interest in biology would be more inclined to participate. As this thesis assumes a context where information contributors are non-experts with respect to the intended information uses by project sponsors (in this case, biologists), the impact of modeling on completeness should be even greater in purely novice populations.

7.5 Chapter Conclusion

This chapter investigates the impact of conceptual modeling on the data completeness dimension of IQ in UGC using a field experiment in the context of citizen science in biology. The empirical evidence demonstrates that users assigned to an implementation derived from class-based conceptual modeling report fewer observations than users assigned to the alternative instance-based condition that follows modeling principles proposed in Chapter 5. Users in the instance-based condition also reported a greater number of new classes of interest. This demonstrates the advantages of modeling UGC using the principles proposed in this thesis over traditional approaches in capturing unanticipated phenomena. Appendix 4 summarizes the findings of the field experiment.

The findings from the field experiment are consistent with those from the laboratory experiments provided in Chapter 4. As in Chapter 4, the field experiment also provides evidence of potential information loss as well as of the prevalence of classes at levels higher-than species (including basic-level categories). Thus, the field experiment provides additional support for Hypothesis H-1.2 (information loss) and also demonstrates the importance of allowing users to report instances at different levels of a

classification hierarchy (which, as demonstrated in Chapter 4 results in higher classification accuracy).

The findings from the field experiment provide empirical evidence for the advantages of the proposed principles of modeling UGC and the proposed definition of crowd IQ. The next chapter considers the contributions of the thesis to the theory and practice of conceptual modeling, IQ and UGC.

8 Contributions, Future Work and Conclusions

User-generated content enables organizations to call upon the collective intelligence of people to support analysis and decision making. Among other uses, contributions of ordinary people expand an organization's "sensor" network, making it possible to collect large amounts of data from highly diverse audiences. Despite the ongoing effort to harness the "wisdom of crowds", unresolved issues of information quality and modeling may significantly curtail adoption of UGC. This thesis provides a theoretical understanding of the nature of information quality and offers theory-based principles to improve crowd IQ. The thesis makes a number of contributions to theory and practice.

8.1 Contributions to Research and Practice

8.1.1 Reconceptualizing IQ

This thesis attempts to open the black box of crowd IQ and argues that important differences exist between traditional organizational settings and crowdsourcing applications. This requires extending the prevailing data consumer focus of IQ definitions, as they ignore the characteristics of crowd (volitional) information creation and do not reflect information contributors' perspectives. A new definition of crowd IQ is proposed: the extent to which stored information represents the phenomena of interest to data consumers (and project sponsors), as perceived by information contributors. This definition explicitly excludes the traditional "fitness for use" conceptualization of IQ.

Rather, it is *use-agnostic*, recognizing that “the phenomena...as perceived by information contributors” accommodates both known uses and future, unanticipated uses.

This thesis provided theoretical arguments and empirical evidence of the advantages of approaching IQ from the contributors' perspectives. These include findings of:

- a) higher accuracy when modeling using classes more natural to data contributors (in UGC settings these are typically basic-level categories) (Chapter 4);
- b) higher accuracy when allowing data contributors to report information freely, without predefined structures (Chapter 4);
- c) higher dataset completeness in an IS that implements instance-based principles of modeling UGC compared with an IS that implements class-based approaches to modeling and focuses on the information needs of the data consumers (i.e., scientists).

These results are novel and provide strong empirical evidence of the advantages of the novel IQ perspective. The contribution of reconceptualizing crowd IQ is in recognizing the pivotal role of information contributors in UGC settings. This recognition leads naturally to a search for more effective designs sensitive to information contributors, while remaining cognizant of the information needs of data consumers.

8.1.2 Exposing Class-based Approaches to Conceptual Modeling as a Factor

Contributing to Poor Crowd IQ

This thesis increases our understanding of the nature of IQ challenges in UGC. Issues of quality in UGC have been receiving increased attention (Alabri and Hunter

2010; Arazy et al. 2011; Liu and Ram 2011; Prestopnik and Crowston 2011; Wiggins et al. 2011). One common assumption is that low quality of UGC is caused by low domain expertise and low levels of motivation of online contributors (see Chapter 2). This thesis contributes to this body of research by demonstrating that in addition to these factors, low crowd IQ may be caused by the approaches to conceptual modeling in the UGC applications. Specifically, the empirical evidence presented in this thesis suggests that traditional approaches to conceptual modeling may have negative impact on accuracy, information loss and dataset completeness dimensions of IQ.

Using three laboratory experiments (Chapter 4) this thesis provides empirical evidence of the negative impact of class-based conceptual models on information accuracy. The results of the experiments demonstrate that accuracy is contingent on the classes used to model a domain. The results show that accuracy in UGC settings decreases when data collection is guided by classes at levels that correspond with predefined uses of data by project sponsors (i.e., biological species). At the same time, accuracy increases when data collection is guided by classes at generic levels. This finding suggests the potential benefit in identifying and modeling UGC applications using generic-level classes. This thesis further suggests cognitive psychology as a theoretical reference for identification of such classes (i.e., basic-level categories).

At the same time, using these generic classes, however, undermines information completeness (causing information loss). This thesis proposed a novel dimension of IQ, information loss (Chapter 3). Following theories of ontology and cognition, I argue that using classes to store information about instances will always result in a failure to fully

capture reality, no matter how “good” the chosen classes are. As (ontologically) classes are unable to capture all instance attributes that might be observed, class-based conceptual models will result in information loss as long as contributors are able to observe attributes of an instance not implied by the class(es) they can provide.

The empirical evidence for the potential prevalence of information loss in UGC was provided in Chapter 4. In Experiment 1, non-expert participants provided significantly more low-level, specific, attributes than more generic attributes. Additional evidence for information loss was obtained in the field experiment (Chapter 7), where users of the instance-based version of NLNature provided attributes that offered additional information not inferable from the classification labels attached to instances. The proposition that class-based models engender information loss (i.e., Proposition 2 in Chapter 3) implies that potentially valuable information may be routinely lost in existing class-based UGC applications.

Another limitation of classification structures is demonstrated in Experiment 3 (Chapter 4) that compared unconstrained (free-form) and schema-mediated data collection. This comparison shows that accuracy does not necessarily improve when intuitive and accurate options are provided for users. The results of Experiment 3 demonstrated that the overall classification accuracy in the free-form data collection condition was significantly greater than in either single or multi-level conditions. This further indicates the potential consequences of using a class-based conceptual modeling: while predefined classes provide cues that may guide users to correct choices, they may also bias users to wrong classification decisions.

Finally, class-based conceptual modeling may have negative impact on dataset completeness. In the field experiment (Chapter 7), users assigned to an instantiation based on the traditional class-based conceptual modeling reported fewer species observations compared with the alternative instance-based condition. Users in the instance-based condition also reported a greater number of new classes of interest as well as more instances of new classes. This demonstrates that, by focusing on classes that are useful to organizations, UGC projects may be not capturing all relevant phenomena when classes used to represent these phenomena are incongruent with the views of data contributors.

This thesis demonstrates a connection between conceptual modeling approaches and IQ. Traditionally conceptual modeling and IQ have been considered quite different domains. Conceptual modeling research explored effective domain representations (Mylopoulos 1998, Olivé 2007, Parsons and Wand 2008, Wand and Weber 2002), while IQ research examined data accuracy, completeness, and fitness for use in already designed systems (Lee et al. 2006, Pipino et al. 2002, Tayi and Ballou 1998, Wang and Strong 1996). Novel IQ challenges in user-generated datasets illustrate a critical role for conceptual modeling in information quality, which is likely to be applicable in internal corporate settings as well as in the environment of UGC.

This thesis is one of the first attempts to establish *theoretical* antecedents of information quality dimensions (Wand and Wang 1996; Wang and Strong 1996) and discover mechanisms for improving quality. Despite extensive research on, and the centrality of IQ to organizational decision making, relatively little is known about what

causes low quality data - resulting in what has been called "a significant gap in the IS research" (Petter et al. 2013, p. 30).

By showing specific ways conceptual modeling affect IQ, this thesis demonstrates the importance of conducting conceptual modeling and IQ research in tandem and calls for greater consideration of IQ in future conceptual modeling research and practice. The novel connections between conceptual modeling and IQ should make it easier to more effectively leverage conceptual modeling in improving IQ. Likewise, a better understanding of IQ implications promises to inform conceptual modeling theory and practice and suggest directions for improving modeling methods and grammars.

8.1.3 Novel Approaches to Improving IQ

This research points to the potential of an alternative data structure, based on attributes and instances, to improve crowd IQ. By allowing instances to exist independent of any classification, an application does not *a priori* constrain the potential information that can be stored. Thus, contributors can supply attributes based on their levels of domain expertise without having to pass a (potentially incorrect) classification judgement. Such an approach assumes neither a particular use of the data nor a minimal level of domain expertise and is, in that sense, use- and expertise-agnostic.

This thesis further contributes by providing a "proof by construction" and demonstrates the application of the proposed principles of modeling UGC by re-designing a real IS, NLNature (www.nlnature.com). The NLNature design attests to the feasibility of the proposed principles and also provides a blueprint that practitioners can follow when developing UGC projects.

This research demonstrates a context in which instance-and-attribute based data collection and storage can lead to higher quality information for those who benefit from UGC. The approach is clearly useful when contributors lack domain knowledge or do not share the conceptual models (class structures) of information consumers. Additionally, where there is the opportunity to capture a diverse range of instance information (attributes that would not be expressed in a shared conceptual model), an instance-and-attribute approach offers flexibility that cannot be achieved using a predetermined class structure. Such flexibility is likely to be valuable when there is a reasonable prospect of using information for purposes other than those envisioned when a system was designed. It can be combined with a traditional class-based approach (which might also include basic level classes) when there is a range from novice to expert contributors, who can be identified when contributions are reported. In addition, experts who classify at a fine level can also be given opportunities to report additional attribute information.

8.2 Future Research

This thesis provides a basis for a significant future research program that builds on the theoretical arguments and empirical findings presented here. Key directions for future research are provided below.

8.2.1 Impact of Conceptual Modeling on Other IQ Dimensions

This thesis provides a theoretical argument and empirical evidence for the impact of conceptual modeling on central the IQ dimension of accuracy and completeness. One avenue for future research is extending the theoretical understanding of the relationship

between modeling and crowd IQ by investigating other IQ dimensions. IQ research recognizes several dozen dimensions including consistency, timeliness, believability, accessibility, security, value-added, ease of manipulation, and freedom from error (Lee et al. 2002; Wang and Strong 1996). For example, *data believability* (i.e., whether a decision maker believes this data is correct, complete or current) becomes particularly important when dealing with UGC as the context of data creation and even the identity of data contributors maybe unknown. Employing an instance-based approach to modeling UGC should promote confidence in the crowd data once decision makers become aware that the contributors were not constrained and biased by potentially incongruent conceptual structures. Future work may provide additional guidance for employing UGC in organizational decision making by increasing understanding of other dimensions of crowd IQ.

8.2.2 Impact of Contributor-oriented IQ on Data Consumers

The prevailing conceptualization of IQ as 'fitness for use' explicitly guided IS towards ways to serve the needs of the organization. This thesis demonstrated a number of advantages of an alternative perspective in IQ that focuses on information contributors. An important question that remains open is the impact of the contributor-oriented IQ on data consumers. Here, one issue is whether organizations can take advantage of the novel affordances of contributor-oriented IQ. For example, data that is more faithful to the crowd's perspective can be leveraged in designing better customer-facing products or services or redesigning internal processes to make them more agile and flexible (see Kharabe and Lyytinen 2013). Future research can also examine challenges that data

consumers (e.g., scientists) may face when interpreting and analyzing instance-based data as well as opportunity this data presents.

8.2.3 From UGC to Other Domains

Another area for future research is applying the proposed perspective on IQ in other domains. Although this work is framed in terms of UGC, it can also be applied to traditional corporate systems when information about entities might be used for purposes not anticipated when a system was designed. For example, if the sole purpose of an asset management system is to keep track of accounting information about assets, a traditional class-based structure might be adequate. If, however, it is discovered that the performance of assets depends on the conditions under which they are used, but this relationship was not anticipated when the (class-based) asset management system was designed, the system would need to be redesigned to capture additional attributes of assets reflecting the conditions of use (entailing a detailed analysis of the kinds of conditions that matter and the specific impact on attributes of assets). In contrast, an instance-based system would be able to accommodate additional attributes of specific assets independent of any classification. Such an approach can help in generating new ways of conceptualizing phenomena in a seemingly familiar and well-understood domain.

Many enterprise-wide and inter-organizational IS integrate large and often heterogeneous views of data (Vitale and Johnson 1988, Zhu and Wu 2011). Much like the UGC setting explored in this thesis, such integration creates the possibility of under-representing the perspectives of many individual data contributors. As Kent (1978) noted: "we can share a common enough view of [reality] for most of our working purposes, so

that reality does appear to be objective and stable... But the chances of achieving such a shared view become poorer when we try to encompass broader purposes, and to involve more people" (p. 203).

Future work can investigate the applicability and advantages of use-agnostic IQ and instance-based modeling in more traditional, corporate settings.

8.2.4 Development of an Instance-based Conceptual Modeling Grammar

An interesting question for future research is whether development following the proposed modeling principles can be further enhanced with the help of conceptual modeling scripts. The case of NLNature provided an example of converting the proposed principles into a real IS. In the scenario provided, the analysis phase essentially proceeds without relying on modeling scripts, such as Entity-Relationship diagrams. The principles proposed in Chapter 5 may guide development of conceptual modeling grammars - or rules and constructs (Burton-Jones et al. 2009; Gemino and Wand 2004) that analysts can use to create "instance-based" conceptual modeling scripts. The principles can both suggest ways to extend existing grammars as well as guide the development of new ones.

While prevailing conceptual modeling grammars are driven by abstraction-based representations, they already contain the constructs proposed in Chapter 5. Specifically instances (things) have been used in conceptual modeling under *similar* terms of *object*, *entity* or *instance* (Chen 1976; Evermann and Wand 2005; Parsons and Wand 1997). Similarly, many modeling grammars contain the notion of classes, attributes and relationship types (for review, see Hull and King 1987; Peckham and Maryanski 1988). This means that the proposed principles can be used to extend the existing grammars to

take advantage of the familiarity of analysts with the notations for representing these constructs, as well as the capabilities of the existing visual modeling software. At the same time, popular grammars such as ER and UML are founded on the principles of representation by abstraction, which fundamentally differs from the instance-based representation advanced here. Another option may involve extending grammars that lack graphical components, but share some properties with the proposed principles in this thesis such as those based on the Entity–attribute–value model or prolog / datalog (Patel-Schneider and Horrocks 2007). Future work can investigate whether and how existing modeling grammars can be modified to be more congruent with the principles proposed here.

An alternative to re-using existing grammars is to develop a new conceptual modeling grammar. This permits creating a grammar that is more faithful to the proposed principles. For example, analysts may survey a sample of users and create models of a sample of instances and attributes. Such conceptual models would represent concrete instances rather than abstractions. While this means that these models are fundamentally incomplete, analyzing these attributes provides an early glimpse into what the actual data would look like, supports communication during development and guide design choices (e.g., whether or not to limit attributes to a predefined list).

8.2.5 Addressing Challenges to Instance-and-attribute Approaches

Notwithstanding the advantages of the instance-based approach to crowd IQ demonstrated in this thesis, it has a number of challenges that can be addressed in future studies. One is managing a large number of attributes. As with classes, attributes of

interest may not all be known at the time a system is designed. With potentially a very large set of attributes, it is necessary to devise mechanisms to guide contributors to select from available attributes. This may necessitate grouping attributes in some way, thus negating some of the potential benefits of an instance-based model.

Another issue when allowing contributors to report attributes in a relatively unconstrained manner is standardizing data to make it amenable to analysis. In particular, when users are free to specify attributes, heterogeneity in reporting is likely to result in observations with (slightly) different names for semantically equivalent attributes (synonymy). This limitation can be addressed at both the input and post-processing levels. On input, it is possible to guide contributors to attributes by displaying potential matches for partially specified attributes and allowing contributors to select from them (without constraining users to these options). One area for future research is to examine the effectiveness of techniques for standardizing instance-based data.

There is also a concern about the effort expended in providing a large number of attributes. Free-form attribute collection may become difficult to use as it would excessive entail typing - this may be especially concerning for small mobile keypads. There seems to be a need for novel approaches in data collection interfaces that could be more faithful to the proposed modeling principles. A promising future direction involves developing hybrid conversational data entry interfaces that allow users to type or speak the attributes and classes. Some advantages to such IS include lessening of the typing burden and greater accessibility (especially on wearable and miniaturized devices) (see also Shneiderman 2000). Another strategy in support of instance-based user input is

automatic attribute extraction. In this case attributes are generated without direct human effort. Some attributes can originate in sensor data provided by the browsing agents. A common practice on the internet is to fetch browser-supplied data (e.g., IP address, screen size, resolution); cookies are also widely used to store and exchange information. Future extensions of NLNature may exploit these technologies to gain a better understanding about a user and the operating environment (this information can then assist in interpreting the attribute-data provided by the user). A system can also leverage any additional information that the user-operating agent provides. Thus when NLNature is accessed via a location-aware device (e.g., smart phone or smart wearable), the geo-coordinates of the instance can be extracted automatically. This can also include date and time of the sighting, temperature, humidity, wind speed and other environmental indicators, without asking users directly for this data. Similarly, if users provide photos or videos of an instance, automatic feature detection and extraction algorithms (Hsu et al. 2002; e.g., Nixon and Aguado 2012) may be employed (and, optionally, the features they generate could be provided to users for validation). The approaches suggested above open a wide avenue for future research.

8.2.6 Combining Instance-based Modeling with Traditional Modeling

An important area for future investigation is modeling under a hybrid abstraction-based/instance-based approach. In practice most IS are likely to be on different points on the development continuum, as some aspects of a system could remain relatively fixed and amenable to abstraction-driven modeling. For example legal, security and reporting considerations could be embedded in software consistent with some fixed convention

rather than left open to judgment of individual users. Similarly, a requirement to exchange data with legacy systems may suggest pre-specifying some structures in advance (Atzeni et al. 2013). This raises questions about how to integrate the proposed modeling principles with traditional abstraction-driven modeling. Currently little is known about these issues and much scope exists in learning how to strike a balance between different modeling approaches.

8.3 Thesis Conclusions

As organizations invite diverse and unpredictable user-generated content into the world of internal decision making, they face the challenge of managing the quality of such datasets. Applications like citizen science create opportunities to collect and analyze data in ways that are not otherwise possible. Despite the potential for online engagement with citizen science and online users in general, the prevailing assumptions and practices underlying data collection in these projects may limit the amount of relevant information that organizations are able to harness.

The online environment in which user contributions are being made is different from the traditional internal corporate environment of data management in three important ways that affect information quality. First, within a controlled environment it is possible to ensure a high level of data input quality (via training, input controls and other measures). In contrast, in projects harnessing user input the organization often has little control over the domain expertise and motivation of potential contributors. Second, in a corporate environment, databases are generally initially designed with specific applications and uses in mind, making it possible to tailor the design of the database using

a set of domain classes that are well-understood within the organization. In contrast, the potential uses of UGC may not be fully known when the system is designed and deployed. Finally, traditional design assumes the success of information systems is contingent on how well such systems capture and implement user requirements (Appan and Browne 2010). Users' views of reality are central to seminal IQ conceptualizations (Wand and Wang 1996, Wang and Strong 1996). In many UGC projects (such as those in citizen science) with a distributed, diverse, and potentially uncommitted user base, the traditional process of information requirements determination is practically unachievable.

This research focuses attention on the black box of crowd IQ. By evaluating existing practices against theories of philosophy and human cognition, this thesis draws attention to a number of critical questions and provides insights on how crowd information quality can be conceptualized and improved.

Bibliography

- Abiteboul, S. 1997. "Querying semi-structured data," *Database Theory—ICDT'97*, pp. 1-18.
- Alabri, A. and Hunter, J. 2010. "Enhancing the Quality and Trust of Citizen Science Data," *IEEE Sixth International Conference on e-Science*, pp. 81-88.
- Allen, G. N. and March, S. T. 2012. "A Research Note on Representing Part-Whole Relations in Conceptual Modeling," *MIS Quarterly*, (36:3), pp. 945-964.
- Ambler, S. 2003. *Agile database techniques: effective strategies for the agile software developer*, Wiley, Hoboken, NJ.
- Andriole, S. J. 2010. "Business Impact of Web 2.0 Technologies," *Communications of the ACM*, (53:12), pp. 67-79.
- Angles, R. and Gutierrez, C. 2008. "Survey of graph database models," *ACM Computing Surveys*, (40:1), pp. 1-39.
- Anwar, S. and Parsons, J. 2010. "An ontological foundation for agile modeling with UML," *Americas Conference on Information Systems*, pp. 1-12.
- Appan, R. and Browne, G. J. 2012. "The Impact of Analyst-Induced Misinformation on the Requirements Elicitation Process," *MIS Quarterly*, (36:1), pp. 85-106.
- Appan, R. and Browne, G. J. 2010. "Investigating Retrieval-Induced Forgetting During Information Requirements Determination," *Journal of the Association for Information Systems*, (11:5), pp. 250-275.
- Arazy, O. and Kopak, R. 2011. "On the measurability of information quality," *Journal of the American Society for Information Science and Technology*, (62:1), pp. 89-99.
- Arazy, O., Kumar, N. and Shapira, B. 2010. "A theory-driven design framework for social recommender systems," *Journal of the Association for Information Systems*, (11:9), pp. 455-490.
- Arazy, O., Nov, O., Patterson, R. and Yeo, L. 2011. "Information Quality in Wikipedia: The Effects of Group Composition and Task Conflict," *Journal of Management Information Systems*, (27:4), pp. 71-98.
- Atzeni, P., Jensen, C. S., Orsi, G., Ram, S., Tanca, L. and Torlone, R. 2013. "The relational model is dead, SQL is dead, and I don't feel so good myself," *ACM SIGMOD Record*, (42:1), pp. 64-68.
- Ballou, D. P. and Pazer, H. L. 1995. "Designing Information Systems to Optimize the Accuracy-timeliness Tradeoff," *Information Systems Research*, (6:1), pp. 51.
- Ballou, D. P. and Pazer, H. L. 1985. "Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems," *Management Science*, (31:2), pp. 150-162.

- Ballou, D. P., Wang, R., Pazer, H. and Tayi, G. K. 1998. "Modeling Information Manufacturing Systems to Determine Information Product Quality," *Management Science*, (44:4), pp. 462-484.
- Bargh, J. A. and Chartrand, T. L. 1999. "The unbearable automaticity of being." *American Psychologist*, (54:7), pp. 462-479.
- Barsalou, L. W. 1983. "Ad hoc categories," *Memory & Cognition*, (11), pp. 211-227.
- Barwise, P. and Meehan, S. 2010. "The One Thing You Must Get Right When Building a Brand," *Harvard Business Review*, (88:12), pp. 80-84.
- Batini, C., Lenzerini, M. and Navathe, S. B. 1986. "A comparative analysis of methodologies for database schema integration," *Computing Surveys*, (18:4), pp. 323-364.
- Batini, C., Cappiello, C., Francalanci, C. and Maurino, A. 2009. "Methodologies for data quality assessment and improvement," *Computing Surveys*, (41:3), pp. 1-52.
- Berlin, B., Breedlove, D. E. and Raven, P. H. 1973. "General Principles of Classification and Nomenclature in Folk Biology," *American Anthropologist*, (75:1), pp. 214-242.
- Bishr, M. and Mantelas, L. 2008. "A trust and reputation model for filtering and classifying knowledge about urban growth," *GeoJournal*, (72:3), pp. 229-237.
- Boakes, E. H., McGowan, P. J., Fuller, R. A., Chang-qing, D., Clark, N. E., O'Connor, K. and Mace, G. M. 2010. "Distorted views of biodiversity: spatial and temporal bias in species occurrence data," *PLoS Biology*, (8-6), e1000385.
- Bodart, F., Patel, A., Sim, M. and Weber, R. 2001. "Should Optional Properties Be Used in Conceptual Modelling? A Theory and Three Empirical Tests," *Information Systems Research*, (12:4), pp. 384-405.
- Bonter, D. N. and Cooper, C. B. 2012. "Data validation in citizen science: a case study from Project FeederWatch," *Frontiers in Ecology and the Environment*, (10:6), pp. 305-307.
- Bowker, G. C. and Star, S. L. 2000. *Sorting things out: classification and its consequences*, MIT Press, Cambridge, MA.
- Braun, S., Schmidt, A., Walter, A., Nagypal, G. and Zacharias, V. 2007. "Ontology Maturing: a Collaborative Web 2.0 Approach to Ontology Engineering," *16th International World Wide Web Conference WWW2007*.
- Browne, G. J. and Parsons, J. 2012. "More Enduring Questions in Cognitive IS Research," *Journal of the Association for Information Systems*, (13:12), pp. 1000-1011.
- Browne, G. J. and Ramesh, V. 2002. "Improving information requirements determination: a cognitive perspective," *Information & Management*, (39:8), pp. 625-645.

- Bunge, M. 1977. *Treatise on basic philosophy: Ontology I: the furniture of the world*, Reidel, Boston, MA.
- Burton-Jones, A., Clarke, R., Lazarenko, K. and Weber, R. 2013. "Is Use of Optional Attributes and Associations in Conceptual Modeling Always Problematic? Theory and Empirical Tests," *International Conference on Information Systems*, pp. 1-14.
- Burton-Jones, A., Wand, Y. and Weber, R. 2009. "Guidelines for Empirical Evaluations of Conceptual Modeling Grammars," *Journal of the Association for Information Systems*, (10:6), pp. 495-532.
- Burton-Jones, A. and Meso, P. N. 2006. "Conceptualizing Systems for Understanding: An Empirical Test of Decomposition Principles in Object-Oriented Analysis," *Information Systems Research*, (17:1), pp. 38-60.
- Burton-Jones, A. and Meso, P. N. 2008. "The Effects of Decomposition Quality and Multiple Forms of Information on Novices' Understanding of a Domain from a Conceptual Model," *Journal of the Association for Information Systems*, (9:12), pp. 748-802.
- Burton-Jones, A. and Weber, R. 1999. "Understanding relationships with attributes in entity-relationship diagrams," *20th International Conference on Information Systems*, pp. 214-228.
- Carey, S. 2009. *The Origin of Concepts*, Oxford University Press, New York, USA.
- Cattell, R. 2011. "Scalable SQL and NoSQL data stores," *ACM SIGMOD Record*, (39:4), pp. 12-27.
- Cha, M., Kwak, H., Rodriguez, P., Ahn, Y. and Moon, S. 2007. "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," pp. 1-14.
- Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., Chandra, T., Fikes, A. and Gruber, R. E. 2008. "Bigtable: A distributed storage system for structured data," *ACM Transactions on Computer Systems*, (26:2), pp. 4-23.
- Chau, M. and Xu, J. 2012. "Business Intelligence in Blogs: Understanding Consumer Interactions and Communities." *MIS Quarterly*, (36:4), pp. 1189-1216.
- Checkland, P. and Holwell, S. 1998. *Information, systems, and information systems: making sense of the field*, John Wiley & Sons, Inc, Hoboken, NJ.
- Chen, H., Chiang, R. H. and Storey, V. C. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact." *MIS Quarterly*, (36:4), pp. 1165-1188.
- Chen, P. 1976. "The entity-relationship model - toward a unified view of data," *ACM Transactions on Database Systems*, (1:1), pp. 9-36.
- Chen, P. 2006. "Suggested Research Directions for a New Frontier – Active Conceptual Modeling," *Conceptual Modeling: ER'2006*, pp. 1-4.

- Choudhury, V. 1997. "Strategic choices in the development of interorganizational information systems," *Information Systems Research*, (8:1), pp. 1-24.
- Christen, P. 2012. "A survey of indexing techniques for scalable record linkage and deduplication," *IEEE Transactions on Knowledge and Data Engineering*, (24:9), pp. 1537-1555.
- Clarke, R., Burton-Jones, A. and Weber, R. 2013. "Improving the Semantics of Conceptual-Modeling Grammars: A New Perspective on an Old Problem," *International Conference on Information Systems*, pp. 1-17.
- Coleman, D. J., Georgiadou, Y. and Labonte, J. 2009. "Volunteered Geographic Information: The Nature and Motivation of Producers," *International Journal of Spatial Data Infrastructures Research*, (4:1), pp. 332-358.
- Collins, H., R. Evans. 2007. *Rethinking Expertise*. University of Chicago Press, Chicago, IL.
- Corter, J. and Gluck, M. 1992. "Explaining basic categories: Feature predictability and information," *Psychological Bulletin*, (111:2), pp. 291-303.
- Cruse, D. A. 1977. "The Pragmatics of Lexical Specificity," *Journal of Linguistics*, (13:2), pp. 153-164.
- Culnan, M. J., McHugh, P. J. and Zubillaga, J. I. 2010. "How large U.S. companies can use Twitter and other social media to gain business value." *MIS Quarterly Executive*, (9:4), pp. 243-259.
- Daugherty, T., Eastin, M. and Bright, L. 2008. "Exploring Consumer Motivations for Creating User-Generated Content," *Journal of Interactive Advertising*, (8:2), pp. 16-25.
- DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., Sivasubramanian, S., Voshall, P. and Vogels, W. 2007. "Dynamo: amazon's highly available key-value store," *ACM SIGOPS Operating Systems Review*, pp. 205-220.
- DeLone, W. H. and McLean, E. R. 1992. "Information Systems Success: The Quest for the Dependent Variable," *Information Systems Research*, (3:1), pp. 60-95.
- Delort, J., Arunasalam, B. and Paris, C. 2011. "Automatic Moderation of Online Discussion Sites," *International Journal of Electronic Commerce*, (15:3), pp. 9-30.
- Dickinson, J. L., Zuckerberg, B. and Bonter, D. N. 2010. "Citizen science as an ecological research tool: challenges and benefits," *Annual Review of Ecology, Evolution, and Systematics*, (41:1), pp. 112-149.
- Doan, A. and Halevy, A. Y. 2005. "Semantic-integration research in the database community - A brief survey," *Ai Magazine*, (26:1), pp. 83-94.

- Doan, A., Ramakrishnan, R. and Halevy, A. Y. 2011. "Crowdsourcing systems on the World-Wide Web," *Communications of the ACM*, (54:4), pp. 86-96.
- Dobing, B. and Parsons, J. 2006. "How UML is used," *Communications of the ACM*, (49:5), pp. 109-113.
- Easterby-Smith, M., Thorpe, R. and Jackson, P. R. 2012. *Management research*, SAGE, Los Angeles, CA.
- Eden, C. and Ackermann, F. 1998. *Making Strategy: The Journey of Strategic Management*, Sage Publications.
- Eimas, P. and Quinn, P. 1994. "Studies on the Formation of Perceptually Based Basic-Level Categories in Young Infants," *Child Development*, (65:3), pp. 903-917.
- Erickson, L., Petrick, I. and Trauth, E. 2012. "Hanging with the right crowd: Matching crowdsourcing need to crowd characteristics," *AMCIS 2012 Proceedings*, pp. 1-9.
- Estes, W. K. 1996. *Classification and Cognition*, Oxford University Press, Oxford, UK.
- Evermann, J. 2008. "Theories of meaning in schema matching: A review." *Journal of Database Management*, (19:3), pp. 55-82.
- Evermann, J. and Wand, Y. 2006. "Ontological modeling rules for UML: An empirical assessment," *Journal of Computer Information Systems*, (46), pp. 14-29.
- Evermann, J. and Wand, Y. 2005. "Ontology based object-oriented domain modelling: fundamental concepts," *Requirements Engineering*, (10:2), pp. 146-160.
- Falkowski, A. and Feret, B. 1990. "Prototype and exemplar models in categorization: A simulatory comparative analysis," *Polish Psychological Bulletin*, (21:3), pp. 199-211.
- Fan, W. and Geerts, F. 2012. "Foundations of Data Quality Management," *Synthesis Lectures on Data Management*, (4:5), pp. 1-217.
- Figl, K. and Derntl, M. 2011. "The impact of perceived cognitive effectiveness on perceived usefulness of visual conceptual modeling languages," *Conceptual Modeling-ER 2011*, pp. 78-91.
- Flanagin, A. and Metzger, M. 2008. "The credibility of volunteered geographic information," *GeoJournal*, (72:3), pp. 137-148.
- Fodor, J. A. 1998. *Concepts: where cognitive science went wrong*, Clarendon Press, Oxford, UK.
- Fortson, L., Masters, K., Nichol, R., Borne, K., Edmondson, E., Lintott, C., Raddick, J., Schawinski, K. and Wallin, J. 2011. "Galaxy Zoo: Morphological Classification and Citizen Science," *Advances in Machine Learning and Data Mining for Astronomy*, pp. 1-11.
- Foster-Smith, J. and Evans, S. M. 2003. "The value of marine ecological data collected by volunteers," *Biological Conservation*, (113:2), pp. 199-213.

- Franklin, M. J., Kossmann, D., Kraska, T., Ramesh, S. and Xin, R. 2011. "CrowdDB: answering queries with crowdsourcing," *ACM SIGMOD International Conference on Management of Data*, pp. 61-72.
- Gallagher, K., J. Parsons, K. D. Foster. 2001. A tale of two studies: Replicating 'Advertising effectiveness and content evaluation in print and on the web.'. *Journal of Advertising Research* **41** (4) 71-81.
- Gallaugh, J. and Ransbotham, S. 2010. "Social Media and Customer Dialog Management at Starbucks," *MIS Quarterly Executive*, (9:4), pp. 197-212.
- Galloway, A. W. E., Tudor, M. T. and Haegen, W. M. V. 2006. "The Reliability of Citizen Science: A Case Study of Oregon White Oak Stand Surveys," *Wildlife Society Bulletin*, (34:5), pp. 1425-1429.
- Gangi, P. M. D., Wasko, M. and Hooker, R. 2010. "Getting Customers' Ideas to Work for You: Learning from Dell how to Succeed with Online User Innovation Communities," *MIS Quarterly Executive*, (9:4), pp. 163-178.
- Gao, G., McCullough, J. S., Agarwal, R. and Jha, A. K. 2010. "Are doctors created equal? An investigation of online ratings by patients," *Workshop on Information Systems Economics*, St. Louis, MO.
- Gemino, A. and Wand, Y. 2005. "Complexity and clarity in conceptual modeling: comparison of mandatory and optional properties," *Data & Knowledge Engineering*, (55:3), pp. 301-326.
- Gemino, A. and Wand, Y. 2004. "A framework for empirical evaluation of conceptual modeling techniques," *Requirements Engineering*, (9:4), pp. 248-260.
- Gentner, D. and Markman, A. B. 1997. "Structure Mapping in Analogy and Similarity," *American Psychologist*, (52:1), pp. 45-56.
- Ghose, A., Goldfarb, A. and Han, S. 2012. "How is the Mobile Internet Different? Search Costs and Local Activities," *Information Systems Research*, pp. 1-19.
- Gibbons, J. and Chakraborti, S. 1992. *Nonparametric Statistical Inference. Third Edition*, Marcel Dekker, Inc., New York, NY.
- Girres, J. and Touya, G. 2010. "Quality Assessment of the French OpenStreetMap Dataset," *Transactions in GIS*, (14:4), pp. 435-459.
- Gleasure, B., Feller, J. and O'Flaherty, B. 2012. "Procedurally Transparent Design Science Research: A Design Process Model," *International Conference on Information Systems - ICIS 2012*, pp. 1-19.
- Goldstone, R. L. and Medin, D. L. 1994. "Similarity, Interactive Activation, and Mapping: An Overview", Keith J. Holyoak and John A. Barnden (eds.), in *Analogical Connections*. Ablex Publishing, Westport, CT US.

- Goldwater, M. B., Tomlinson, M. T., Echols, C. H. and Love, B. C. 2011. "Structural Priming as Structure-Mapping: Children Use Analogies From Previous Utterances to Guide Sentence Production," *Cognitive Science*, (35:1), pp. 156-170.
- Good, P. I. 2001. *Resampling Methods: A Practical Guide to Data Analysis. Second Edition*, Springer, New York - Berlin.
- Goodchild, M. 2007. "Citizens as sensors: the world of volunteered geography," *GeoJournal*, (69:4), pp. 211-221.
- Gould, J. D. and Lewis, C. 1985. "Designing for usability: key principles and what designers think," *Communications of the ACM*, (28:3), pp. 300-311.
- Gregor, S. and Jones, D. 2007. "The Anatomy of Design Theory," *Journal of the Association for Information Systems*, (8:5), pp. 312-335.
- Grossman, M., Aronson, J. E. and McCarthy, R. V. 2005. "Does UML make the grade? Insights from the software development community," *Information and Software Technology*, (47:6), pp. 383-397.
- Guizzardi, G. 2010. "Theoretical foundations and engineering tools for building ontologies as reference conceptual models," *Semantic Web*, (1:1), pp. 3-10.
- Guizzardi, G. and Halpin, T. 2008. "Ontological foundations for conceptual modelling," *Appl. Ontol.*, (3:1-2), pp. 1-12.
- Hahn, U., Chater, N. and Richardson, L. B. 2003. "Similarity as transformation," *Cognition*, (87:1), pp. 1-32.
- Haklay, M. and Weber, P. 2008. "OpenStreetMap: User-Generated Street Maps," *IEEE Pervasive Computing*, (7:4), pp. 12-18.
- Hamel, N. J., Burger, A. E., Charleton, K., Davidson, P., Lee, S., Bertram, D. F. and Parrish, J. K. 2009. "Bycatch and beached birds: Assessing mortality impacts in coastal net fisheries using marine bird strandings," *Marine Ornithology*, pp. 41-60.
- Hand, E. 2010. "People power," *Nature*, (466:7307), pp. 685-687.
- Hanna, R., Rohm, A. and Crittenden, V. L. 2011. "We're all connected: The power of the social media ecosystem," *Business Horizons*, (54:3), pp. 265-273.
- Harnad, S. 2005. "To Cognize is to Categorize: Cognition is Categorization", H. Cohen and C. Lefebvre (eds.), in *Handbook of Categorization in Cognitive Science*, Elsevier Science, Amsterdam.
- Harnad, S. R. 1990. *Categorical Perception: The Groundwork of Cognition*, Cambridge University Press, Cambridge, MA.
- Hayes, A. F. 1996. "Permutation test is not distribution-free: Testing $H_0: \rho = 0$." *Psychological Methods*, (1:2), pp. 184-198.

- Heath, T. and Bizer, C. 2011. *Linked Data: Evolving the Web Into a Global Data Space*, Morgan & Claypool Publishers, San Rafael, CA.
- Heipke, C. 2010. "Crowdsourcing geospatial data," *ISPRS Journal of Photogrammetry and Remote Sensing*, (65:6), pp. 550-557.
- Hemsley, J. and Mason, R. M. 2012. "The Nature of Knowledge in the Social Media Age: Implications for Knowledge Management Models," pp. 3928-3937.
- Hevner, A., March, S., Park, J. and Ram, S. 2004. "Design science in information systems research," *MIS Quarterly*, (28:1), pp. 75-105.
- Hirschheim, R., Klein, H. K. and Lyytinen, K. 1995. *Information Systems Development and Data Modeling: Conceptual and Philosophical Foundations*, Cambridge University Press, Cambridge.
- Hochachka, W. M., Fink, D., Hutchinson, R. A., Sheldon, D., Wong, W. and Kelling, S. 2012. "Data-intensive science applied to broad-scale citizen science," *Trends in Ecology & Evolution*, (27:2), pp. 130-137.
- Holyoak, K. J. and Koh, K. 1987. "Surface and Structural Similarity in Analogical Transfer," *Memory & Cognition*, (15:4), pp. 332-340.
- Hsu, R., Abdel-Mottaleb, M. and Jain, A. K. 2002. "Face detection in color images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (24:5), pp. 696-706.
- Hull, R. and King, R. 1987. "Semantic database modeling: survey, applications, and research issues," *ACM Comput. Surv.*, (19:3), pp. 201-260.
- Imai, S. 1977. "Pattern similarity and cognitive transformations," *Acta Psychologica*, (41:6), pp. 433-447.
- Indulska, M., Hovorka, D. S. and Recker, J. 2011. "Quantitative approaches to content analysis: identifying conceptual drift across publication outlets," *European Journal of Information Systems*, (21:1), pp. 49-69.
- Jacobson, I., Booch, G., Rumbaugh, J., Rumbaugh, J. and Booch, G. 1999. *The unified software development process*, Addison-Wesley, Reading MA.
- Johnson, P. A. and Sieber, R. E. 2012. "Situating the Adoption of VGI by Government", Daniel Sui, Sarah Elwood and Michael Goodchild (eds.), in *Crowdsourcing Geographic Knowledge*, Springer, Netherlands.
- Jolicoeur, P., Gluck, M. A. and Kosslyn, S. M. 1984. "Pictures and names: Making the connection," *Cognitive Psychology*, (16:2), pp. 243-275.
- Jones, R. A. and Rosenberg, S. 1974. "Structural representations of naturalistic descriptions of personality," *Multivariate Behavioral Research*, (9:2), pp. 217-230.
- Juran, J. M. and Gryna, F. M. 1988. *Juran's quality control handbook*, McGraw-Hill.

- Jussim, L., Nelson, T. E., Manis, M. and Soffin, S. 1995. "Prejudice, stereotypes, and labeling effects: Sources of bias in person perception." *Journal of Personality and Social Psychology*, (68:2), pp. 228-246.
- Kahn, B. K., Strong, D. M. and Wang, R. Y. 2002. "Information quality benchmarks: product and service performance," *Communications of the ACM*, (45:4), pp. 184-192.
- Kahneman, D. D. 1992. "The reviewing of object files: Object-specific integration of information," *Cognitive Psychology*, (24:2), pp. 175-219.
- Kaldor, N. 1961. "Capital Accumulation and Economic Growth", F. A. Lutz and D. C. Hague (eds.), in *The Theory of Capital*, Macmillan, London.
- Kallio, S. 2012. "An assessment of the reliability of online volunteer-based bird observation networks," *Memorial Univesrity Biology Undergraduate Research Symposium*, Memorial Univesrity, St. John's, Canada.
- Kent, W. 1978. *Data and reality: basic assumptions in data processing reconsidered*, North-Holland Pub. Co., Amsterdam.
- Kharabe, A. and Lyytinen, K. J. 2013. "Is Implementing ERP Like Pouring Concrete Into a Company? Impact of Enterprise Systems on Organizational Agility," *International Conference on Information Systems*, pp. 1-20.
- Kim, S., Robson, C., Zimmerman, T., Pierce, J. and Haber, E. M. 2011. "Creek watch: pairing usefulness and usability for successful citizen science," pp. 2125-2134.
- Kimura, T., Gillett, W. D. and Cox, J. 1985. "A design of a data model based on abstraction of symbols," *The Computer Journal*, (28:3), pp. 298-308.
- Kittur, A., Chi, E., Pendleton, B., Sun, B. and Mytkowicz, T. 2007. "Power of the few vs wisdom of the crowd: Wikipedia and the rise of the bourgeoisie," *World wide web*, pp. 1-9.
- Klibanoff, R. S. and Waxman, S. R. 2000. "Basic Level Object Categories Support the Acquisition of Novel Adjectives: Evidence from Preschool-Aged Children," *Child Development*, (71:3), pp. 649-659.
- Kluge, J., Kargl, F. and Weber, M. 2007. "The Effects of the Ajax Technology on Web Application Usability." *International Conference on Web Information Systems and Technologies*, pp. 289-294.
- Korpela, E. J. 2012. "SETI@ home, BOINC, and volunteer distributed computing," *Annual Review of Earth and Planetary Sciences*, (40), pp. 69-87.
- Kotiadis, K. and Robinson, S. 2008. "Conceptual modelling: knowledge acquisition and model abstraction," pp. 951-958.
- Krogstie, J., Lyytinen, K., Opdahl, A., Pernici, B., Siau, K. and Smolander, K. 2003. "ER/IFIP8. 1 Workshop on Conceptual Modelling Approaches to Mobile Information Systems Development (MobIMod 2002)-Mobile Information

- Systems--Research Challenges on the Conceptual and Logical Level," *Lecture Notes in Computer Science*, (2784), pp. 124-135.
- Krumm, J., Davies, N. and Narayanaswami, C. 2008. "User-Generated Content," *IEEE Pervasive Computing*, (7:4), pp. 10-11.
- Kuechler, W. and Vaishnavi, V. 2012. "A Framework for Theory Development in Design Science Research: Multiple Perspectives." *Journal of the Association for Information Systems*, (13:6), pp. 395-423.
- Kwak, H., Lee, C., Park, H. and Moon, S. 2010. "What is Twitter, a social network or a news media?" pp. 591-600.
- Lakoff, G. 1987. *Women, fire, and dangerous things : what categories reveal about the mind*, University of Chicago Press, Chicago.
- Lambert, N. M., Graham, S. M. and Fincham, F. D. 2009. "A Prototype Analysis of Gratitude: Varieties of Gratitude Experiences," *Personality and Social Psychology Bulletin*, (35:9), pp. 1193-1207.
- Landis, J. R. and Koch, G. G. 1977. "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, (33:1), pp. 159-174.
- Lassaline, M. E., Wisniewski, E. J. and Medin, D. L. 1992. "Basic Levels in Artificial and Natural Categories: Are all Basic Levels Created Equal?", Burns Barbara (ed.), in *Advances in Psychology*, North-Holland.
- Lee, A. S. and Baskerville, R. L. 2003. "Generalizing generalizability in information systems research," *Information Systems Research*, (14:3), pp. 221-243.
- Lee, Y. W. and Strong, D. M. 2003. "Knowing-why about data processes and data quality," *Journal of Management Information Systems*, (20:3), pp. 13-39.
- Lee, Y. W., Strong, D. M., Kahn, B. K. and Wang, R. Y. 2002. "AIMQ: A methodology for information quality assessment," *Information & Management*, (40:2), pp. 133-146.
- Lee, Y. W. 2006. *Journey to data quality*, MIT Press, Cambridge, MA.
- Lee, Y. W. 2003. "Crafting Rules: Context-Reflective Data Quality Problem Solving," *Journal of Management Information Systems*, (20:3), pp. 93-119.
- Lintott, C. J., Schawinski, K., Keel, W., Van Arkel, H., Bennert, N., Edmondson, E., Thomas, D., Smith, D. J. B., Herbert, P. D., Jarvis, M. J., Virani, S., Andreescu, D., Bamford, S. P., Land, K., Murray, P., Nichol, R. C., Raddick, M. J., Slosar, A., Szalay, A. and Vandenberg, J. 2009. "Galaxy Zoo: Hanny's Voorwerp, a quasar light echo?" *Monthly Notices of the Royal Astronomical Society*, (399:1), pp. 129-140.
- Liu, J. and Ram, S. 2011. "Who does what: Collaboration patterns in the wikipedia and their impact on data quality," *ACM Transactions on Management Information Systems*, pp. 175-180.

- Liu, C., Chrysanthis, P. K. and Chang, S. 1994. "Database Schema Evolution through the Specification and Maintenance of Changes on Entities and Relationships", Pericles Loucopoulos (ed.), in *Entity-Relationship Approach—ER'94 Business Modelling and Re-Engineering*, Springer, Berlin.
- Liu, X., Lu, M., Ooi, B. C., Shen, Y., Wu, S. and Zhang, M. 2012. "CDAS: A crowdsourcing data analytics system," *VLDB Endowment*, (5:10), pp. 1040-1051.
- Lohr, S. 2012. "The age of big data", *New York Times*, February 12, 2012.
- Louv, R., Dickinson, J. L. and Bonney, R. 2012. *Citizen Science: Public Participation in Environmental Research*, Cornell University Press, Ithaca, NY.
- Lukyanenko, R. and Parsons, J. 2013. "Is traditional conceptual modeling becoming obsolete?" *Conceptual Modeling: ER'2013*, pp. 1-14.
- Lukyanenko, R. and Evermann, J. 2011. "A survey of cognitive theories to support data integration," *AMCIS 2011 Proceedings*, pp. 1-15.
- Lukyanenko, R. and Parsons, J. 2012. "Conceptual modeling principles for crowdsourcing," *Proceedings of the 1st International Workshop on Multimodal Crowd Sensing*, pp. 3-6.
- Ma, Z. M. and Yan, L. 2008. "A Literature Overview of Fuzzy Database Modeling," *Journal of Information Science and Engineering*, (24:1), pp. 189-202.
- Mackechnie, C., Maskell, L., Norton, L. and Roy, D. 2011. "The role of 'Big Society' in monitoring the state of the natural environment," *Journal of Environmental Monitoring*, (13:10), pp. 2687-2691.
- Madnick, S. E., Wang, R. Y., Lee, Y. W. and Zhu, H. 2009. "Overview and Framework for Data and Information Quality Research," *Journal of Data and Information Quality*, (1:1), pp. 1-22.
- Majchrzak, A. N. N. and More, P. H. B. 2011. "Emergency! Web 2.0 to the Rescue!" *Communications of the ACM*, (54:4), pp. 125-132.
- March, S. T. and Smith, G. F. 1995. "Design and natural science research on information technology," *Decision Support Systems*, (15:4), pp. 251-266.
- March, S. and Allen, G. 2012. "Toward a social ontology for conceptual modeling," *11th Symposium on Research in Systems Analysis and Design*, pp. 57-62.
- Markman, A. B. and Wisniewski, E. J. 1997. "Similar and different: The differentiation of basic-level categories," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, (23:1), pp. 54-70.
- Markus, M. L., Steinfield, C. W. and Wigand, R. T. 2006. "Industry-wide information systems standardization as collective action: the case of the US residential mortgage industry," *MIS Quarterly*, pp. 439-465.
- Martinez, W. L., Martinez, A. and Solka, J. 2004. *Exploratory Data Analysis with MATLAB*, Taylor & Francis, New York, NY.

- Mason, R. O. and Mitroff, I. I. 1973. "A program for research on management information systems," *Management Science*, (19:5), pp. 475-487.
- Masri, K. 2009. "Conceptual model design for better understanding", *PhD Thesis*, Simon Frazer University.
- Mayden, R. L. 2002. "On biological species, species concepts and individuation in the natural world," *Fish and Fisheries*, (3:3), pp. 171-196.
- McCloskey, M. and Glucksberg, S. 1978. "Natural categories: Well defined or fuzzy sets?" *Memory & Cognition*, (6:4), pp. 462-472.
- McGinnes, S. 2011. "Conceptual Modelling for Web Information Systems: What Semantics can be Shared?", Olga De Troyer, Claudia Bauzer Medeiros, Roland Billen, Pierre Hallot, Alkis Simitsis and Hans Van Mingroot (eds.), Springer, Berlin / Heidelberg.
- Medin, D. L. and Schaffer, M. M. 1978. "Context theory of classification learning," *Psychological Review*, (85:3), pp. 207-238.
- Mervis, C. B. and Crisafi, M. A. 1982. "Order of Acquisition of Subordinate-, Basic-, and Superordinate-Level Categories," *Child Development*, (53:1), pp. 258-266.
- Michael, L. M., Isabel, G., Javid, S. and Thomas, J. P. 2008. "Object detection and basic-level categorization: Sometimes you know it is there before you know what it is," *Psychonomic Bulletin & Review*, (15:1), pp. 28-35.
- Mix, K. S. 2008. "Surface similarity and label knowledge impact early numerical comparisons," *British Journal of Developmental Psychology*, (26:1), pp. 13-32.
- Moody, D. L. 2005. "Theoretical and practical issues in evaluating the quality of conceptual models: current state and future directions," *Data & Knowledge Engineering*, (55:3), pp. 243-276.
- Murphy, G. L. 2004. *The big book of concepts*, MIT Press, Cambridge, Mass.
- Murphy, G. L. and Wisniewski, E. J. 1989. "Categorizing Objects in Isolation and in Scenes - What a Superordinate Is Good For," *Journal of Experimental Psychology-Learning Memory and Cognition*, (15:4), pp. 572-586.
- Murphy, G. L. 1982. "Cue validity and levels of categorization," *Psychological Bulletin*, (91:1), pp. 174-177.
- Mylopoulos, J. 1992. "Conceptual Modeling and Telos", P. Loucopoulos and R. Zicari (eds.), in *Conceptual Modeling, Databases, and CASE: An Integrated View of Information Systems Development*, John Wiley & Sons, Inc., New York, NY.
- Mylopoulos, J. 1998. "Information modeling in the time of the revolution," *Information Systems*, (23:3-4), pp. 127-155.
- Mylopoulos, J. and Borgida, A. 2006. "Properties of Information Modeling Techniques for Information Systems Engineering", Peter Bernus, Kai Mertins and Günter

- Schmidt (eds.), in *Handbook on Architectures of Information Systems*, Springer, Berlin Heidelberg.
- Nelson, R. R., Todd, P. A. and Wixom, B. H. 2005. "Antecedents of information and system quality: an empirical examination within the context of data warehousing," *Journal of Management Information Systems*, (21:4), pp. 199-235.
- Newell, A. and Card, S. K. 1985. "The prospects for psychological science in human-computer interaction," *Human-Computer Interaction*, (1:3), pp. 209-242.
- Nixon, M. S. and Aguado, A. S. 2012. *Feature Extraction & Image Processing for Computer Vision*, Academic Press, Waltham, MA.
- Nosofsky, R. M. 1986. "Attention, Similarity, and the Identification-Categorization Relationship," *Journal of Experimental Psychology: General*, (115:1), pp. 39-57.
- Nov, O., Arazy, O. and Anderson, D. 2011a. "Dusting for science: motivation and participation of digital citizen science volunteers," *2011 iConference*, pp. 68-74.
- Nov, O., Arazy, O. and Anderson, D. 2011b. "Technology-Mediated Citizen Science Participation: A Motivational Model," *ICWSM*, pp. 1-8.
- Nunamaker, J. F., Chen, M. and Purdin, T. D. 1991. "Systems development in information systems research," *Journal of Management Information Systems*, (7:3), pp. 89-106.
- Olivé, A. 2007. *Conceptual modeling of information systems*, Springer, Berlin Heidelberg New York.
- Parfitt, I. 2013. "Citizen science in conservation biology: best practices in the geoweb era," *Masters Thesis*, University of British Columbia.
- Parnas, D. L. 1972. "A technique for software module specification with examples," *Communications of the ACM*, (15:5), pp. 330-336.
- Parsons, J. and Wand, Y. 2013. "Cognitive Principles to Support Information Requirements Agility," *Advanced Information Systems Engineering Workshops*, pp. 192-197.
- Parsons, J., Lukyanenko, R. and Wiersma, Y. 2011. "Easier citizen science is better," *Nature*, (471:7336), pp. 37-37.
- Parsons, J. and Wand, Y. 2008. "Using cognitive principles to guide classification in information systems modeling," *MIS Quarterly*, (32:4), pp. 839-868.
- Parsons, J. 2003. "Effects of Local Versus Global Schema Diagrams on Verification and Communication in Conceptual Data Modeling," *Journal of Management Information Systems*, (19:3), pp. 155-184.
- Parsons, J. 2011. "An Experimental Study of the Effects of Representing Property Precedence on the Comprehension of Conceptual Schemas," *Journal of the Association for Information Systems*, (12:6), pp. 441-462.

- Parsons, J. 1996. "An Information Model Based on Classification Theory," *Management Science*, (42:10), pp. 1437-1453.
- Parsons, J. and Cole, L. 2005. "What do the pictures mean? Guidelines for experimental evaluation of representation fidelity in diagrammatical conceptual modeling techniques," *Data & Knowledge Engineering*, (55:3), pp. 327-342.
- Parsons, J. and Wand, Y. 1997. "Choosing classes in conceptual modeling," *Communications of the ACM*, (40:6), pp. 63-69.
- Parsons, J. and Wand, Y. 2000. "Emancipating Instances from the Tyranny of Classes in Information Modeling," *ACM Transactions on Database Systems*, (25:2), pp. 228–268.
- Parsons, J. and Wand, Y. 1997. "Using objects for systems analysis," *Communications of the ACM*, (40:12), pp. 104-110.
- Parssian, A., Sarkar, S. and Jacob, V. S. 2004. "Assessing Data Quality for Information Products: Impact of Selection, Projection, and Cartesian Product," *Management Science*, (50:7), pp. 967-982.
- Patel-Schneider, P. F. and Horrocks, I. 2007. "A comparison of two modelling paradigms in the Semantic Web," *Web Semantics: Science, Services and Agents on the World Wide Web*, (5:4), pp. 240-250.
- Peckham, J. and Maryanski, F. 1988. "Semantic data models," *ACM Computing Surveys*, (20:3), pp. 153-189.
- Petter, S., DeLone, W. and McLean, E. R. 2013. "Information Systems Success: The Quest for the Independent Variables," *Journal of Management Information Systems*, (29:4), pp. 7-62.
- Pipino, L. L., Lee, Y. W. and Wang, R. Y. 2002. "Data quality assessment," *Communications of the ACM*, (45:4), pp. 211-218.
- Piskorski, M. J. 2011. "Social Strategies That Work." *Harvard Business Review*, (89:11), pp. 116.
- Pohl, K. 1994. "The three dimensions of requirements engineering: a framework and its applications," *Information Systems*, (19:3), pp. 243-258.
- Pokorny, J. 2013. "NoSQL databases: a step to database scalability in web environment," *International Journal of Web Information Systems*, (9:1), pp. 69-82.
- Posner, M. I. 1993. *Foundations of cognitive science*, MIT Press, MIT.
- Prestopnik, N. R. and Crowston, K. 2011. "Gaming for (Citizen) Science: Exploring Motivation and Data Quality in the Context of Crowdsourced Science through the Design and Evaluation of a Social-Computational System," "Computing for Citizen Science" Workshop at the IEEE eScience Conference, pp. 1-28.

- Provost, F. and Fawcett, T. 2013. *Data Science for Business: What you need to know about data mining and data-analytic thinking*, O'Reilly Media, Inc., Sebastopol, CA.
- Recker, J., Michael Rosemann, Green, P. and Indulska, M. 2011. "Do ontological deficiencies in modeling grammars matter?" *MIS Quarterly*, (35:1), pp. 57-79.
- Redman, T. C. 1996. *Data quality for the information age*, Artech House, Norwood, MA.
- Reeves, C. A. and Bednar, D. A. 1994. "Defining Quality: Alternatives and Implications," *The Academy of Management Review*, (19:3), pp. 419-445.
- Rhemtulla, M. and Hall, D. 2009. "Basic-level kinds and object persistence," *Memory & Cognition*, (37:3), pp. 292-301.
- Rips, L. J., Blok, S. and Newman, G. 2006. "Tracing the identity of objects," *Psychological Review*, (113:1), pp. 1-30.
- Robal, T., Haav, H. and Kalja, A. 2007. "Making Web Users' Domain Models Explicit by Applying Ontologies", Jean-Luc Hainaut (ed.), in *Advances in Conceptual Modeling-Foundations and Applications*, Springer, Berlin / Heidelberg.
- Robson, C., Hearst, M., Kau, C. and Pierce, J. 2013. "Comparing the use of social networking and traditional media channels for promoting citizen science," *Computer Supported Cooperative Work (CSCW)*, pp. 1463-1468.
- Rosch, E. 1974. "Basic-Level Objects in Natural Categories," *Bulletin of the Psychonomic Society*, (4:Na4), pp. 246-246.
- Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M. and Boyesbraem, P. 1976. "Basic Objects in Natural Categories," *Cognitive Psychology*, (8:3), pp. 382-439.
- Rosch, E. and Muller, S. 1978. "Judgments Based on Classification Theory - Restrictions in Classification of Stereotypes," *Zeitschrift Fur Sozialpsychologie*, (9:3), pp. 246-256.
- Rosch, E. 1978. "Principles of Categorization", Eleanor Rosch and Barbara Lloyd (eds.), in *Cognition and Categorization*, John Wiley & Sons Inc, Hoboken, NJ.
- Rosenberg, S. and Sedlak, A. 1972. "Structural Representations of Implicit Personality Theory", Leonard Berkowitz (ed.), in *Advances in Experimental Social Psychology*, Academic Press, .
- Roussopoulos, N. and Karagiannis, D. 2009. "Conceptual Modeling: Past, Present and the Continuum of the Future", Alexander Borgida, Chaudhri V, Paolo Giorgini and Eric Yu (eds.), in *Conceptual Modeling: Foundations and Applications*, Springer, Berlin / Heidelberg.
- Saghafi, A. and Wand, Y. 2014. "Conceptual Models? A Meta-Analysis of Empirical Work," *Hawaii International Conference on System Sciences*, pp. 1-14.

- Samuel, B. 2012. "Reconceptualizing Conceptual Schema Comprehension: Understanding the Role of Instantiation and Abstraction," *10th Symposium on Research in Systems Analysis and Design*, pp. 21-27.
- Scholl, B. J. 2002. *Objects and Attention*, MIT Press, Cambridge, MA.
- Sein, M., Henfridsson, O., Purao, S., Rossi, M. and Lindgren, R. 2011. "Action design research " *MIS Quarterly*, (35:1), pp. 37-56.
- Shanks, G., Tansley, E., Nuredini, J., Tobin, D. and Weber, R. 2008. "Representing part-whole relations in conceptual modeling: an empirical evaluation," *MIS Quarterly*, (32:3), pp. 553-573.
- Shepard, R. 1962. "The analysis of proximities: Multidimensional scaling with an unknown distance function. I," *Psychometrika*, (27:2), pp. 125-140.
- Sheppard, S., Wiggins, A. and Terveen, L. 2014. "Capturing Quality: Retaining Provenance for Curated Volunteer Monitoring Data," *Computer Supported Cooperative Work and Social Computing*, pp. 1-17.
- Sherman, R. 2007. "The Trial-and-Error Method for Data Integration," *DM Review*, (17:2), pp. 31-31.
- Shneiderman, B. 2000. "The limits of speech recognition," *Communications of the ACM*, (43:9), pp. 63-65.
- Sieber, R. 2012. "Participatory Geospatial Web 2.0: Theory Meets Practice," *GIScience Biennial Conference*, pp. 1-3.
- Sieber, R. 2006. "Public participation geographic information systems: A literature review and framework," *Annals of the Association of American Geographers*, (96:3), pp. 491-507.
- Silvertown, J. 2009. "A new dawn for citizen science," *Trends in Ecology & Evolution*, (24:9), pp. 467-471.
- Silvertown, J. 2010. "Taxonomy: include social networking," *Nature*, (467:7317), pp. 788-788.
- Sim, J., C. C. Wright. 2005. "The kappa statistic in reliability studies: use, interpretation, and sample size requirements." *Physical therapy* (85:3), pp. 257-268.
- Smith, E. E. and Medin, D. L. 1981. *Categories and concepts*, Harvard University Press, Cambridge, Mass.
- Smith, J. M. and Smith, D. C. P. 1977. "Database abstractions: aggregation and generalization," *ACM Transactions on Database Systems*, (2:2), pp. 105-133.
- Smith, L. B. 2005. "Emerging Ideas about Categories", L. Gershkoff-Stowe and D. H. Rakison (eds.), in *Building Object Categories in Developmental Time*, L. Erlbaum Associates, Mahwah, NJ.



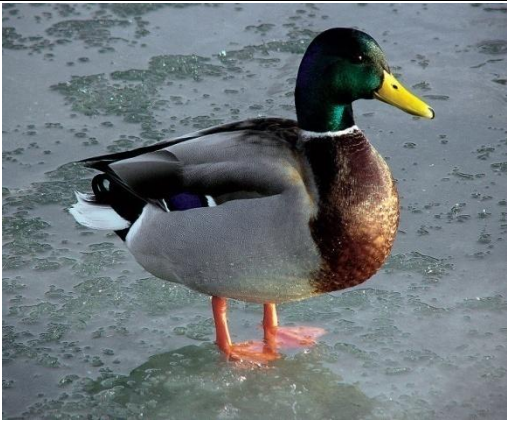

- Snäll, T., Kindvall, O., Nilsson, J. and Pärt, T. 2011. "Evaluating citizen-based presence data for bird monitoring," *Biological Conservation*, (144:2), pp. 804.
- Spaccapietra, S. and Parent, C. 1994. "View integration: A step forward in solving structural conflicts," *IEEE Transactions on Knowledge and Data Engineering*, (6:2), pp. 258-274.
- Stoller, J. 2009. "Data quality: Familiar problem, new challenges," *CMA Management*, (83:7), pp. 37-38.
- Storey, V. C., Dewan, R. M. and Freimer, M. 2012. "Data quality: Setting organizational policies," *Decision Support Systems*, (54:1), pp. 434-442.
- Strong, D. M., Lee, Y. W. and Wang, R. Y. 1997. "Data quality in context," *Communications of the ACM*, (40:5), pp. 103-110.
- Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D. and Kelling, S. 2009. "eBird: A citizen-based bird observation network in the biological sciences," *Biological Conservation*, (142:10), pp. 2282-2292.
- Surowiecki, J. 2005. *The wisdom of crowds*, Anchor Books, New York, NY.
- Susarla, A., Oh, J. and Tan, Y. 2012. "Social Networks and the Diffusion of User-Generated Content: Evidence from YouTube," *Information Systems Research*, (23:1), pp. 23-41.
- Tanaka, J. W. and Taylor, M. 1991. "Object categories and expertise: Is the basic level in the eye of the beholder?" *Cognitive Psychology*, (23:3), pp. 457-482.
- Tayi, G. K. and Ballou, D. P. 1998. "Examining data quality," *Communications of the ACM*, (41:2), pp. 54-57.
- Teorey, T. J., Yang, D. and Fry, J. P. 1986. "A logical design methodology for relational databases using the extended entity-relationship model," *ACM Computing Surveys*, (18:2), pp. 197-222.
- Topi, H. and Ramesh, V. 2002. "Human Factors Research on Data Modeling: A Review of Prior Research, An Extended Framework and Future Research Directions", *Journal of Database Management*, (13:2), pp. 3-19.
- Tsichritzis, D. C. and Lochovsky, F. H. 1982. *Data models*, Prentice-Hall, Englewood Cliffs, N.J.
- Tversky, A. 1977. "Features of similarity," *Psychological Review*, (84:4), pp. 327-352.
- Tversky, A. and Gati, I. 1982. "Similarity, separability, and the triangle inequality," *Psychological Review*, (89:2), pp. 123-154.
- Van Kleek, M. G., Styke, W., Schraefel, M. and Karger, D. 2011. "Finders/Keepers: A Longitudinal Study of People Managing Information Scraps in a Micro-note Tool," *SIGCHI Conference on Human Factors in Computing Systems*, pp. 2907-2916.

- Venkatesh, V., Morris, M. G., Davis, G. B. and Davis, F. D. 2003. "User acceptance of information technology: Toward a unified view," *MIS Quarterly*, pp. 425-478.
- Vitale, M. R. and Johnson, H. 1988. "Creating competitive advantage with interorganizational information systems," *MIS Quarterly*, (12:2), pp. 152-165.
- Volkoff, O., Strong, D. M. and Elmes, M. B. 2007. "Technological embeddedness and organizational change," *Organization Science*, (18:5), pp. 832-848.
- Walsham, G. 1993. *Interpreting information systems in organizations*, Wiley Pub.
- Wand, Y. 1996. "Ontology as a foundation for meta-modelling and method engineering," *Information and Software Technology*, (38:4), pp. 281-287.
- Wand, Y. and Weber, R. 2008. "On the deep structure of information systems," *Information Systems Journal*, (5:3), pp. 203-223.
- Wand, Y. and Weber, R. 1995. "On the Deep-Structure of Information-Systems," *Information Systems Journal*, (5:3), pp. 203-223.
- Wand, Y. and Weber, R. 1993. "On the ontological expressiveness of information systems analysis and design grammars," *Information Systems Journal*, (3:4), pp. 217-237.
- Wand, Y. and Weber, R. 2002. "Research commentary: Information systems and conceptual modeling - A research agenda," *Information Systems Research*, (13:4), pp. 363-376.
- Wand, Y. and Weber, R. 1990. "Toward a theory of the deep structure of information systems," *International Conference on Information Systems*, pp. 61-71.
- Wand, Y., Monarchi, D. E., Parsons, J. and Woo, C. C. 1995. "Theoretical foundations for conceptual modelling in information systems development," *Decision Support Systems*, (15:4), pp. 285-304.
- Wand, Y. and Wang, R. Y. 1996. "Anchoring data quality dimensions in ontological foundations," *Communications of the ACM*, (39:11), pp. 86-95.
- Wand, Y. and Weber, R. 2006. "On Ontological Foundations of Conceptual Modeling: A Response to Wyssusek," *Scandinavian Journal of Information Systems*, (18:1), pp. 1-11.
- Wang, R. Y. 1998. "A product perspective on total data quality management," *Communications of the ACM*, (41:2), pp. 58-65.
- Wang, R. Y. and Strong, D. M. 1996. "Beyond accuracy: what data quality means to data consumers," *Journal of Management Information Systems*, (12:4), pp. 5-33.
- Weber, R. 1996. "Are attributes entities? A study of database designers' memory structures," *Information Systems Research*, (7:2), pp. 137-162.
- Wiersma, Y. F. 2010. "Birding 2.0: citizen science and effective monitoring in the Web 2.0 world," *Avian Conservation and Ecology*, (5:2), pp. 13.

- Wiggins, A., Bonney, R., Graham, E., Henderson, S., Kelling, S., LeBuhn, G., Litauer, R., Lots, K., Michener, W. and Newman, G. 2013. "Data management guide for public participation in scientific research", *DataOne Working Group*, pp. 1-41.
- Wiggins, A., Newman, G., Stevenson, R. D. and Crowston, K. 2011. "Mechanisms for Data Quality and Validation in Citizen Science," *IEEE e-Science Workshops (eScienceW)*, 2011 IEEE Seventh International Conference, pp. 14-19.
- Winograd, T. and Flores, F. 1987. *Understanding computers and cognition: a new foundation for design*, Addison-Wesley.
- Wyssusek, B. 2006. "On Ontological Foundations of Conceptual Modelling." *Scandinavian Journal of Information Systems*, (18:1), pp. 63-80.
- Zhang, C., Xie, J., Xie, J., Wu, M., Huang, Y. and Huang, X. 2013. "Detecting the core network of microblog using snowball sampling," *Wireless Personal Multimedia Communications*, pp. 1-5.
- Zhu, H. and Wu, H. 2011. "Quality of data standards: framework and illustration using XBRL taxonomy and instances," *Electronic Markets*, (21:2), pp. 129-139.

Appendix 1: Images Used in Laboratory Experiments in Chapter 4

Images source: Wikimedia Commons; The order as appeared in one of the experimental sessions.

	
Killer whale (<i>Orcinus orca</i>)	Old man's beard (<i>Usnia</i> spp.)
	
Mallard duck (<i>Anas platyrhynchos</i>)	Eastern Coyote (<i>Canis latrans</i>)



Calypso orchid (*Calypso bulbosa*)



Caspian tern (*Hydroprogne caspia*)



Red squirrel (*Tamiasciurus hudsonicus*)









Moose (*Alces alces*)



Blue winged teal (*Anas discors*)



Labrador tea (*Ledum groenlandicum*)

	
<p>Indian pipe (<i>Monotropa uniflora</i>)</p>	<p>Common tern (<i>Sterna hirundo</i>)</p>
	
<p>Fireweed (<i>Epilobium angustifolium</i>)</p>	<p>Eastern chipmunk (<i>Tamias striatus</i>)</p>
	
<p>American robin <i>Turdus migratorius</i></p>	<p>Sheep laurel (<i>Kalmia angustifolia</i>)</p>



False morel (*Gyromitra esculenta*)



Greater yellowlegs (*Tringa melanoleuca*)



Caribou (*Rangifer tarandus*)



Red Fox (*Vulpes vulpes*)



Blue jay (*Cyanocitta cristata*)



Atlantic Salmon (*Salmo salar*)



Lung lichen (*Lobaria pulmonaria*)



Spotted sandpiper (*Actitis macularia*)

Appendix 2: Summary of Options Provided in Experiments 2 and 3 of Chapter 4

Table A2.1. Options provided in Experiment 2, single-level condition (* indicates correct option)

Species	Species-level
Atlantic Salmon	Arctic char, Atlantic cod, Atlantic mackerel, Atlantic salmon*, Brook trout, Conner, Pike, Rainbow trout, Shad
Blue Winged Teal	Blue-winged Teal*, Bufflehead, Common Eider, Common Merganser, Common Teal, Harlequin Duck, Mallard, Northern Pintail, Wood Duck
Calypso Orchid	Calypso Orchid*, Green-fringed Orchid, Indian pipe, Ladyslipper Orchid, Lesser Stitchwort, Northern Bracted Frog Orchid, Pitcher plant, True Forget-me-not, Tuberous Grasspink
Caspian Tern	Arctic Tern, Bonaparte's Gull, Caspian Tern*, Common Tern, Herring Gull, Iceland Gull, Killdeer, Parasitic jaeger, Pomarine jaeger
Common Tern	Arctic Tern, Bonaparte's Gull, Caspian Tern, Common Tern*, Herring Gull, Iceland Gull, Killdeer, Parasitic jaeger, Pomarine jaeger
False morel	Chanterelle, Common morel, False Morel*, Fly agaric, Horse mushroom, Jelly leaf fungus, Larch Boletus, Ornate-stalked Boletus, True Morel
Fireweed	Alpine Campion, Fireweed*, Labrador Tea, Northern Twayblade, Rhodora, Sheep Laurel, Swamp Laurel, Sweet Gale, wild bergamot
Indian Pipe	Calypso Orchid, Indian pipe*, Ladyslipper Orchid, Lesser Stitchwort, Northern Bracted Frog Orchid, Northern Twayblade, Pitcher plant, Rattlesnake Plantain, True Forget-me-not
Mallard duck	American Wigeon, Bufflehead, Common Eider, Common Merganser, Common Teal, Harlequin Duck, Mallard*, Northern Pintail, Wood Duck
Sheep Laurel	Alpine Campion, Fireweed, Labrador Tea, Lesser Stitchwort, Rhodora, Sheep Laurel*, Swamp Laurel, Sweet Gale, True Forget-me-not

Table A2.2. Options provided in Experiment 2, multi-level condition (* indicates correct option)

Species	Basic-level	Species	Subordinate	Superordinate
Atlantic Salmon	Fish*	Atlantic salmon*, Brook trout, Smelt	Diadromous fish*, Ray- finned fish*, Salmon*, Tropical fish	Animal*
Blue Winged Teal	Bird*, Duck*, Goose	Blue-winged Teal*, Wood Duck	Dabbling duck*	Animal*, Warm- blooded organism*, Waterfowl*
Calypso Orchid	Flower*	Calypso Orchid*, Ladyslipper Orchid	Iris, Orchid*	Annual plant, Parasitic plant, Perennial plant*, Plant*
Caspian Tern	Bird*	Caspian Tern*, Herring Gull	Loon, Shorebird, Tern*, Waterfowl	Animal*, Warm- blooded organism*
Common Tern	Bird*	Common Tern*, Iceland Gull	Loon, Shorebird, Tern*, Waterfowl	Animal*, Warm- blooded organism*,
False morel	Mushroom*	Common morel, False Morel*		Ectomycorrhizal fungus, Fungus*, Mycorrhizal fungus*, Plant, Sac fungus*, Saprobe*
Fireweed	Flower*, Shrub	Fireweed*, Sweet Gale	Orchid, Willow-herb*	Annual, Perennial*, Plant*
Indian Pipe	Flower*	Indian pipe*, Ladyslipper Orchid, Pitcher plant		Annual, Fungus, Parasitic plant*, Perennial*, Plant*
Mallard duck	Bird*, Duck*, Goose	Harlequin Duck, Mallard duck*	Dabbling duck*	Animal*, Warm- blooded organism*, Waterfowl*
Sheep Laurel	Flower*, Shrub*	Lesser Stitchwort, Rhodora, Sheep Laurel*	Orchid	Annual, Conifer, Plant*

Table A2.3. Options provided in Experiment 3, single-level condition (* indicates correct option)

Species	Species-level
American Robin	Barn Swallow, Common Grackle, Baltimore Oriole, American Robin*, Evening Grosbeak, House Sparrow, Blue Jay, House Finch, Northern Flicker
Atlantic Salmon	Atlantic salmon*, Rainbow trout, Atlantic mackerel, Brook trout, Pike, Shad, Atlantic cod, Arctic char, Conner
Blue jay	Barn Swallow, Common Grackle, Baltimore Oriole, American Robin, Evening Grosbeak, House Sparrow, Blue Jay*, House Finch, Northern Flicker
Blue Winged Teal	Mallard, Blue-winged Teal*, Common Merganser, King Eider, Bufflehead, Harlequin, Common Eider, Common Teal, Northern Pintail
Calypso Orchid	Calypso Orchid*, Tuberous Grasspink, Pitcher plant, Indian pipe, Lesser Stitchwort, True Forget-me-not, Green-fringed Orchid, Northern Bracted Frog Orchid, Labrador Tea
Caspian Tern	Caspian Tern*, Common Tern, Arctic Tern, Herring Gull, Pomarine jaeger, Killdeer, Parasitic jaeger, Iceland Gull, Bonaparte's Gull
Common Tern	Caspian Tern, Common Tern*, Arctic Tern, Herring Gull, Pomarine jaeger, Killdeer, Parasitic jaeger, Iceland Gull, Bonaparte's Gull
False morel	False Morel*, Larch Boletus, Chanterelle, Common morel, True Morel, King bolete, Ornate-stalked Boletus, Fly agaric, American matsutake
Fireweed	Fireweed*, Sheep Laurel, Alpine Campion, Swamp Laurel, Labrador Tea, Sweet Gale, Northern Twayblade, Wild bergamot, Rhodora
Indian Pipe	Indian pipe*, Northern Twayblade, Pitcher plant, Rattlesnake Plantain, Lesser Stitchwort, True Forget-me-not, Calypso Orchid, Northern Bracted Frog Orchid, Labrador Tea
Killer Whale	Minke Whale, Sperm Whale, Killer whale*, Fin Whale, Harbour Porpoise, Right whale, Spinner dolphin, Sei whale, Sowerby's beaked whale
Mallard duck	Mallard*, Common Eider, Common Merganser, King Eider, Bufflehead, Harlequin, American Wigeon, Common Teal, Northern Pintail
Sheep Laurel	Fireweed, Sheep Laurel*, Alpine Campion, Swamp Laurel, Labrador Tea, Sweet Gale, True Forget-me-not, Lesser Stitchwort, Rhodora

Table A2.4. Options provided in Experiment 3, multi-level condition (* indicates correct option)

Species	Basic-level	Species-level	Subordinate	Superordinate
American Robin	Bird*	Common Grackle, American Robin*, Baltimore Oriole	Shorebird, Non-migratory bird	Animal*, Cold-blooded organism, Warm-blooded organism*
Atlantic Salmon	Fish*	Smelt, Atlantic cod, Atlantic salmon*, Brook trout	Trout, Tropical fish, Ray-finned fish*	Animal*
Blue jay	Bird*	Evening Grosbeak, Common Grackle, Blue Jay*	Shorebird, Non-migratory bird	Animal*, Cold-blooded organism, Warm-blooded organism*
Blue Winged Teal	Bird*, Goose	Harlequin, Northern Pintail, Blue-winged Teal*	Loon, Grebe	Waterfowl*, Animal*
Calypso Orchid	Flower*	Pitcher plant, Calypso Orchid*, Ladyslipper Orchid	Orchid*, Iris	Perennial plant*, Annual, Parasitic plant
Caspian Tern	Bird*	Pomarine jaeger, Caspian Tern*, Herring Gull	Tern*, Seagull, Shorebird, Loon	Animal*
Common Tern	Bird*	Killdeer, Iceland Gull, Common Tern*	Tern*, Shorebird, Seagull, Waterfowl	Animal*
False morel	Mushroom*, Flower	Larch Bolete, False Morel*, Common morel		Fungus*, Plant, Puffball, Decomposer*
Fireweed	Flower*, Shrub	Labrador Tea, Fireweed*, Sweet Gale	Orchid, Willow-herb*	Perennial*, Annual
Indian Pipe	Flower*	Ladyslipper Orchid, Indian pipe*, Pitcher plant	Tulip	Fungus, Perennial*, Annual, Parasitic plant*
Killer Whale	Whale*, Fish, Dolphin	Killer whale*, Harbour Porpoise, Spinner dolphin	Diadromous fish	Animal*, Mammal*
Mallard duck	Bird*, Goose	Bufflehead, Mallard*, Harlequin	Loon, Teal	Waterfowl*, Animal*
Sheep Laurel	Flowering shrub*, Flower*, Shrub*	Lesser Stitchwort, Rhodora, Sheep Laurel*	Orchid	Conifer, Annual

Appendix 3. Additional Analysis of the Experiments 2 and 3

In the comparison of single-level vs. multi-level models in Experiments 2 and 3 (Chapter 4), a reasonable question is whether the fact that the single-level treatment has only one correct option while the multi-level treatment has multiple correct options could favor the multi-level condition if participants chose options at random. Here, I show that such a potential confound was not present in the data.

Experiment 2

Table A3.1 compares expected responses by chance with actual responses in Experiment 2, Multi-level condition. In all but one case (False morel), people provided significantly more basic level responses (flower, fish, duck, bird and mushroom) than would be expected by chance. This shows that, despite the presence of other options, including plausible options at the level deemed basic, participants consistently choose options consistent with theoretical predictions and the free-form responses of Experiment 1. The paucity of responses for False morel can be explained by the fact that this mushroom was atypical of its kind and when other options were available, participants preferred to select those rather than the basic level. Consistent with results from Experiment 3 (below), *mushroom* is not the most common response for False morel - the most common is the superordinate *fungus* provided by 24 participants (which is significantly higher than would be expected by chance).

Table A3.1. Comparing expected responses by chance vs. actual responses in Experiment 2: Multi-level condition

Species	Theoretically predicted basic response	Obtained basic responses	All categorical responses*	Predicted responses expected by chance**	p-value Chi-Square (Yates' correction)
Atlantic Salmon	Fish	12	39	4.33	0.001
Blue Winged Teal	Duck	26	39	4.33	0.000
Calypso Orchid	Flower	14	37	4.11	0.000
Caspian Tern	Bird	21	38	4.22	0.000
Common Tern	Bird	13	39	4.33	0.000
False morel	Mushroom	1	34	3.78	0.241
Fireweed	Flower	16	39	4.33	0.000
Indian Pipe	Flower	10	36	4.00	0.006
Mallard duck	Duck	11	39	4.33	0.003
Sheep Laurel	Flower	23	35	3.89	0.000
Total	10	147	375	41.67	0.000

* Responses of “I don't know” not included. ** Determined by multiplying all categorical responses by the chance of obtaining a theoretically predicted response (e.g., for Common tern the expected response is bird, which has 1/9 chance of being selected if guessing at random; of 39 responses provided this means 4.33 responses “bird” would be expected).

Notably, for Blue Winged Teal, no responses *bird* were given; For Mallard duck only one response was *goose*; no responses *bird* were given; for Fireweed all basic-level responses were *flower* (no responses were *shrub*); for Sheep Laurel all basic-level responses were *flower* (no responses were *shrub*). Indeed, of 148 responses at the basic-level given, 147 (the exception was *goose* by one participant) were the correct basic-level categories predicted based on psychological theory. This yields almost 100% response accuracy at the basic-level and significantly contributes to the increase in accuracy over the single level condition. The responses are further consistent with the obtained results in the free-form Experiment 1 (and, later, Experiment 3).

To provide further evidence that responses in the multi-level condition were consistent with theoretical predictions and accuracy in this condition was not merely due to the greater number of correct responses, I further analyzed the distribution of responses by levels.

Table A3.2 shows the distribution of options by classification levels in Experiment 2, Multi-level condition (with the specific options for each species detailed in Appendix 2). Based on this table I calculate the expected matrix of results if participants were to guess at random. To be conservative I ignored the fact that of several basic-level options, I predicted that only particular one is going to be selected (e.g., when evaluating the expected value for Mallard duck for basic-level classes I included all three options as having equal chance of being selected, even though I do not expect this).

Table A3.2. Distribution of options by classification levels in Experiment 2, Multi-level condition

Species	Basic-level	Species-level (one correct)	Subordinate	Super- ordinate	Grand Total
Atlantic Salmon	1	3	4	1	9
Blue Winged Teal	3	2	1	3	9
Calypso Orchid	1	2	2	4	9
Caspian Tern	1	2	4	2	9
Common Tern	1	2	4	2	9
False morel	1	2	0	6	9
Fireweed	2	2	2	3	9
Indian Pipe	1	3	0	5	9
Mallard duck	3	2	1	3	9
Sheep Laurel	2	3	1	3	9
Total	16	23	19	32	90

I then compare this matrix with the actual distribution of results by classification levels (Table A3.3). Since the main issue is whether participants were selecting *predefined options* at random, I exclude any responses given in the "other" field (where participants were free to provide responses at any taxonomic level irrespective of the options already provided).

Table A3.3. Distribution of responses by classification levels in Experiment 2, Multi-level condition

Species	Basic-level	Species-level	Subordinate	Superordinate	Total	p-value Chi-Square
Atlantic Salmon	12	19	8	0	39	0.000
Blue Winged Teal	26	11	1	0	38	0.000
Calypso Orchid	14	6	8	9	37	0.000
Caspian Tern	21	3	13	1	38	0.000
Common Tern	13	9	17	0	39	0.000
False morel	1	0	0	32	33	0.003
Fireweed	16	5	7	9	37	0.021
Indian Pipe	10	19	0	7	36	0.000
Mallard duck	12	27	0	0	39	0.000
Sheep Laurel	23	4	4	4	35	0.000
Total	148	103	58	62	371	0.000

In all cases, the results differ from what is expected by random guessing. In most cases (as illustrated in detail above), the responses favor the basic level (and more specifically, when more than one basic is provided, the most salient is chosen). It is also clear that despite the large number of options at subordinate and superordinate levels, these levels are chosen sparingly (with the exception of *False morel*). After basic, participants prefer to provide responses at the species-level (103 responses total). *This*

shows that, although there were more correct options provided (e.g., including options at sub- and superordinate levels), participants in the multi-level condition generally did not use these levels.

The analysis of the responses given in Experiment 2 shows that, both overall and individually by species, the distribution of responses deviates from what would be expected by chance. This demonstrates that the greater accuracy in the multi-level condition of Experiment 2 was not merely due to the provision of a greater number of correct options. There were clear patterns in the responses that were consistent with theoretical expectations. Participants were drawn to options they naturally prefer in spite of the presence of other correct options. The presence of choices that were congruent with the participants' view of the world resulted in the higher classification accuracy compared with the single-level condition where such congruent options were not given.

Experiment 3

Experiment 3 included three species that were expected to be familiar to the participants that were omitted from Experiment 2. To test the saliency of the basic-level category I provided 3 plausible options at the basic level for Killer whale - *whale*, *dolphin* and *fish*, with *whale* being the only correct option. I expected that most basic-level responses would be *whale*, but that the majority of responses across levels would be *Killer whale*.

For Fireweed and Sheep laurel, I included two options deemed basic - *flower* and *shrub* - where *shrub* was incorrect for Fireweed, but correct for Sheep laurel. In this case,

I also expected *flower* to be chosen (the correct basic for Fireweed and more salient basic for Sheep laurel).

For False morel, I included a new option at the basic level, *flower* (in addition to *mushroom*). While flower is a salient basic level (as demonstrated by previous Experiments 1 and 2), I did not expect participants to use this option, as it would be incorrect.

For Mallard duck and Blue-winged teal, in Experiment 2 participants had two correct options at the basic level - *bird* and *duck*. In both cases all participants chose *duck* (see Table A3.1). In Experiment 3 I included only one correct basic-level option and made a conservative choice of removing *duck* and including *bird*. Since Experiments 1 and 2 suggested a strong preference for *duck*, it was difficult to predict whether option *bird* would be the preferred one in Experiment 3. Moreover, as evidenced from Experiments 1 and 2, Mallard duck appeared to be a relatively familiar kind of organism. So, it was entirely possible that without the (preferred) option *duck*, participants select more specific options.

Table A3.4 compares expected responses by chance with actual responses in Experiment 3, Multi-level condition, for the schema-congruent group of organisms. Here I expect a higher-than random number of responses at the species level. The results strongly confirm the predictions made in this thesis. In all cases, participants selected significantly more options *American robin*, *Killer whale* and *Blue jay* than would be expected by chance alone. These results demonstrate that despite the presence of other

options, including correct options at the basic level (whale, bird), participants choose specific options agreed with their conceptualizations.

Table A3.4. Expected by chance vs. actual responses in Experiment 3: Multi-level condition for the schema-congruent group

Species	Theoretically predicted response (<i>species level</i>)	Obtained species level responses	All categorical responses*	Predicted responses expected by chance	p-value Chi-Square (Yates' correction)
American Robin	American robin	12	21	2.33	0.000
Blue jay	Blue jay	16	20	2.22	0.000
Killer Whale	Killer whale	19	21	2.33	0.000
Total	3 total	47	62	6.89	0.000

* Responses "I don't know" not included.

Table A3.5 compares expected responses by chance with actual responses in Experiment 3, Multi-level condition, for the schema-incongruent group of organisms. Here I expect a higher-than random number of responses at the basic level. The results confirm the predictions. In all but two cases (False morel and Indian pipe), participants selected significantly more options *bird*, *fish*, *flower* than would be expected by random guessing. Consistent with results from Experiment 2, *mushroom* is not the most common response for False morel - the most common is the superordinate *fungus* provided by 12 participants (which is significantly greater than chance). Indian pipe is also insignificant, which can be explained by the typicality effects as well – Indian pipe (which looks more like fungus) does not look like a typical flower or even a flower at all. No clearly preferred response for Indian pipe emerged.

Table A3.5. Expected by chance vs. actual responses in Experiment 3: Multi-level condition for the schema-incongruent group (excluding Mallard Duck and Blue-winged teal)

Species	Theoretically predicted responses (basic level)	Obtained basic responses	All categorical responses	Predicted responses expected by chance	p-value Chi-Square (Yates' correction)
Atlantic Salmon	Fish	6	21	2.33	0.038
Calypso Orchid	Flower	10	19	2.11	0.000
Caspian Tern	Bird	9	20	2.22	0.000
Common Tern	Bird	7	20	2.22	0.004
False morel	Mushroom	2	14	1.56	0.964
Fireweed	Flower	7	18	2.00	0.001
Indian Pipe	Flower	3	19	2.11	0.789
Sheep Laurel	Flower	12	19	2.11	0.000
Total	8 total	56	150	28.22	0.000

While it was difficult to make predictions for Mallard duck and Blue-winged teal due to the removal of the clearly preferred *duck* option, the results obtained were also not surprising. Specifically, for Mallard 7 people responded with *duck* (provided in the “other” field), 1 person selected *bird* and 11 people selected *Mallard duck*. This demonstrates that, despite the removal of the *duck* option, this seems to be the preferred (basic level) option for those unfamiliar with its specific level - *Mallard duck*. A similar pattern was obtained for Blue-winged teal, where 7 responses were *duck* (provided in the "other" field), 4 responses were *bird*, 2 responses were *goose* (incorrect basic) and 5 responses were *Blue-winged teal*. Interestingly, of the 16 responses provided in the "other" field in Experiment 3: Multi-level condition, 14 were *duck*. *Duck* was the sole "other" response the provided in the Single-level condition. While these numbers are not statistically significant, they suggest two things: 1) *duck* is the salient option for the two organisms used; and 2) when this option is not explicitly provided, participants still

volunteer it as a response. This, however, seems to occur mostly in the multi-level condition, while the exposure to the species-level classes in the single-level condition appears to "break" this natural tendency (it is important to note, participants in all conditions were in the same study session and received the same instructions in which I encouraged them to provide responses not necessarily given in the forms, or select "I don't know" if they did not know the answer).

Table A3.6 provides the distribution of options by classification levels in Experiment 3, Multi-level condition (Appendix 2 shows the actual options for each organism). Based on this data, I calculate the expected matrix of results if participants were to guess at random. To be conservative, I ignored the fact that of several basic-level options I predict that only particular one is going to be selected (e.g., below, when evaluating the expected value for Mallard duck for basic-level classes, I included both options as having equal chance of being selected).

I then compare the expected matrix with the actual distribution of the results by classification levels (Table A3.7 and Table A3.8). Since the main issue is whether participants were selecting predefined options at random, I exclude any responses given in the "other" field (where participants were free to provide responses at any taxonomic level irrespective of the options already provided).

Table A3.6. Distribution of options by classification levels in Experiment 3, Multi-level condition

Species	Basic-level	Species-level (one correct)	Subordinate	Super- ordinate	Total
American Robin	1	3	2	3	9
Atlantic Salmon	1	4	3	1	9
Blue jay	1	3	2	3	9
Blue Winged Teal	2	3	2	2	9
Calypso Orchid	1	3	2	3	9
Caspian Tern	1	3	4	1	9
Common Tern	1	3	4	1	9
False morel	2	3	0	4	9
Fireweed	2	3	2	2	9
Indian Pipe	1	3	1	4	9
Killer Whale	3	3	1	2	9
Mallard duck	2	3	2	2	9
Sheep Laurel	3	3	1	2	9
Total	21	40	26	30	117

In all cases of the schema-congruent group (see Table A3.7) the results are dominated by the species-level responses. Notably, basic is the second largest (selected 9 times), with almost no choices at other levels. This shows that while there were more correct options provided (e.g., including options at sub- and superordinate levels), participants in the multi-level condition were generally not using these levels for schema-congruent species.

Table A3.7. Distribution of responses by classification levels in Experiment 3, Multi-level condition for the schema-congruent group

Species	Basic-level	Species-level	Subordinate	Superordinate	Grand Total
American Robin	4	17	0	0	21
Blue jay	4	16	0	0	20
Killer Whale	1	18	0	1	20
Total	9	51	0	1	61

In most cases for the schema-incongruent group (and overall), the results deviate from what is expected from random choices (Table A3.8). As illustrated in detail above, most responses are at the basic level. More specifically, when more than one basic is provided, the correct one is chosen. It is also clear that, despite the large number of options at subordinate and superordinate levels, these levels are chosen sparingly (with the exception of *False morel*). As in Experiment 2, after basic participants preferred to provide responses at the species-level (61 responses total). I note a few insignificant cases. There are two explanations for this: one is the typicality effect, which explains the result for Indian Pipe; this result is generally consistent with Experiments 1 and 2. The second reason is the fact that I “exaggerated” the expected frequency for basic-level categories by assuming that each had equal chance of being selected (e.g., *goose* and *bird*, *flower* and *shrub*), inflating the expected frequencies for this class. This can explain the result for Fireweed, where 7 of 8 basic-level responses were *flower* (and 1 shrub). If one assumes that *flower* is the expected basic, the result for Fireweed becomes significant as well.

Table A3.8. Distribution of responses by classification levels in Experiment 3, Multi-level condition

Species	Basic-level	Species-level	Subordinate	Superordinate	Total	p-value
Atlantic Salmon	6	11	4	0	21	0.022
Blue Winged Teal	6	6	2	0	14	0.087
Calypso Orchid	10	5	2	1	18	0.000
Caspian Tern	9	5	6	0	20	0.000
Common Tern	7	7	6	0	20	0.004
False morel	2	0	0	12	14	0.015
Fireweed	8		5	1	18	0.067
Indian Pipe	3	11	1	4	19	0.081
Mallard duck	1	11	2	0	14	0.004
Sheep Laurel	16	1	2	0	19	0.000
Total	68	61	30	18	177	0.000

As with Experiment 2, in Experiment 3 one can observe few responses at levels other than basic and species. Indeed, of 18 superordinate results, 12 involved False morel. Despite having a large number of options available at subordinate and superordinate level, participants were generally avoiding these levels.

The analysis of responses given in Experiment 3 demonstrates that, both overall and individually by species (when considering typicality effects where applicable), the distribution of responses significantly deviates from what would be expected by chance. **Thus, the greater accuracy in the multi-level condition of Experiment 3 was not merely due to the provision of a greater number of correct options.** Participants were clearly drawn to options with which they were comfortable (which can be predicted based on theory) and discounted other options. The presence of choices that were congruent with the participants' view of the world resulted in higher classification accuracy

compared with the single-level condition, where such congruent options were not provided.

Appendix 4. Summary of the Theoretical Propositions and Empirical Evidence Obtained

	Experiment (Task)	Independent variable(s)	Dependent variable	Hypothesis Supported?
Proposition 1: Classification Accuracy. Class-based conceptual models result in lower information accuracy (more classification errors) when the classes defined in an information system do not match those familiar to the information contributor.				
	Laboratory Experiment 1 (free-form)	Level of classification (species-genus versus basic)	Classification accuracy	Supported
	Laboratory Experiment 2 (fixed-choice)	Level of classification (single versus multilevel class-based model)	Classification accuracy	Supported
	Laboratory Experiment 3 (fixed-choice)	Level of classification (single versus multilevel class-based model)	Classification accuracy	Supported
	Laboratory Experiment 3 (fixed-choice and free-form)	Free-form versus schema-constrained classification	Classification accuracy	Supported
Proposition 2: Information Loss. Class-based conceptual models result in information loss when the class that a contributor uses to record an instance does not imply some attributes of the instance observed by the contributor.				
	Laboratory Experiment 1 (free-form)	Level of attributes (basic versus sub-basic)	Information loss	Supported
Proposition 3: Dataset Completeness. Class-based conceptual models undermine dataset completeness (resulting in fewer instances stored) when the classes defined in an information system do not match those familiar to the information contributor.				
	Field experiment	Class-based versus instance-based models	Data set completeness (number of observations stored)	Supported
			Data set completeness (number of instances of novel species stored)	Supported